From measurement and uncertainty fields to features: a statistical approach for model validation.

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

Antonios Alexiadis

May 2020

Abstract

Models are extensively used across science and engineering to inform decisions with potential severe socio-economic impacts. A question associated with model-driven decisions is to what extent can these models be considered trustworthy and representative of the real world. Providing an answer to this question is partly achieved by comparing model predictions to measured quantities during a process known as model validation. The comparison is achieved with the aid of a performance measure known as a validation metric and the development of metrics capable of assessing the quality of predictions when compared to full-field measurements has been the epicentre of research.

Compared to traditional methods of validation where model predictions are assessed against point measurements, providing a restricted view of the model's capacity to simulate the real-world process, this research utilizes full-field measurements. These can comprise of measurements captured across the field of an object, such as displacement or deformation, exploiting modern measuring capabilities, or may stem from the numerical post-processing of measurements captured from a network of sensors. Their common feature is that both measurements and predictions are defined over a grid consisting of thousands or millions of datapoints.

Utilizing the information content of these gridded fields can be challenging, especially when measurements and predictions lie on different grids. A solution to this problem is via feature extraction techniques such as orthogonal decomposition using Chebyshev polynomials. These techniques enable high-dimensional, spatial data to be described as a collection of pre-defined spatial features in a lower-dimensional space. What is missing however, is a way to accurately represent the measurement uncertainty accompanying those fields in that space. This was accomplished using approximate Bayesian computation where the measurement's feature vector is repeatedly compared against synthetically generated datasets resulting in a distribution representing the measurement and its uncertainty. This representation allows inferences to be drawn, allowing high volumes of data to be efficiently analysed.

Assessing the accuracy of a prediction using spatial measurements was achieved with two novel methods. In the first, the predicted and measured datasets are reconstructed on the same grid with the aid of Chebyshev polynomials. They are subsequently compared in a pixel-wise manner and the percentage of differences exceeding the range corresponding to the measurement uncertainty is the result. Moreover, pixel-wise comparisons can be utilized to identify the location and magnitude of model-experiment deviations. The second method uses the Mahalanobis distance between the prediction's feature vector and the distribution corresponding to the measurement and its uncertainty. The benefit of the Mahalanobis distance is that it delivers a quantitative measure of the similarity between the two based on their feature vector representation, allowing large amounts of predictions and measurements to be easily assessed.

Examples ranging from engineering to ecology and oceanography have been used to demonstrate the wide applicability of the developed techniques.

Acknowledgements

Numerous difficulties have at certain points made the prospect of submitting a research thesis unlikely. However, the passion, enthusiasm and resilience demonstrated by the people surrounding me acted as a guide and motivation in difficult times, to persevere, keep pushing and never quit; after all the end goal is closer than it may seem. This section is devoted to them.

Towards helping me achieve that end goal I would like to give great merit to my primary supervisor, Prof. Eann Patterson. I consider him a mentor and the discussions we had invaluable. His wisdom and expertise made problems that seemed insurmountable disappear. Even if we disagreed at times, as is natural in any relationship, in the end we managed to find the middle ground and kept pursuing towards the truth. His high standards of rigour and relentless enthusiasm in scientific research along with an impressive work morale and focus will always act as guides in my own work and development.

I also feel the need to express my gratitude to my secondary supervisor, Prof. Scott Ferson. As a practitioner of uncertainty and statistics I consider our first meeting (a black swan event) to be one of the highlights of my academic life. The enthusiasm and interest he showed during our meetings along with his fine sense of humour transformed them into a pleasurable and unique learning experience.

Of course, none of the above would have been possible without the continuous support of my parents Nikos and Maria and my sister Ioanna. They always advise me to pursue my dreams irrespective of how tough things may get. I have discovered that achieving goals along the way is an important factor of happiness. I will be always indebted to them.

Happiness of course does not only stem from academic achievements but from personal ones as well. A great thank you to everyone who surrounded me with your love and support. The friends I made and the ones that I loved will always remind me of the joy that lies in the unexpected.

Finally, I would like to express my gratitude towards the funding bodies of this research which include the Engineering and Physical Sciences Research Council and Airbus. Especially, I would like to thank Eszter Szigeti, Linden Harris and Sanjiv Sharma of Airbus for their feedback and advice during our discussions. Their insight and intuitive understanding of new and complex ideas never seized to impress me. I consider myself lucky to have them throughout the research.

Thank you!

Contents

List of Figures

List of Tables

List of abbreviations

ABC approximate Bayesian computation

ACF autocorrelation function

AIAA American Institute of Aeronautics and Astronautics

ASME American Society of Mechanical Engineers

CCC concordance correlation coefficient

CDF cumulative distribution function

CEN European Committee for Standardization

DIC digital image correlation

DoD Department of Defense

EDF, DF empirical distribution function

ENSO El-Niño Southern Oscillation

ESS effective sample size

FEM finite element model

MAE mean absolute error

MCMC Markov chain Monte Carlo

MD Mahalanobis distance

MPIT multivariate probability integral transform

MSE mean squared error

NOAA National Oceanic and Atmospheric Administration

ONI Oceanic Niño Index

PCA principal component analysis

PC principal component

PIT probability integral transform

PM probabilistic metric

RMSE root-mean-square error

ROI region of interest

SD shape descriptor

SG strain gauge

SSIM structural similarity index

STD standard deviation

V&V verification and validation

Nomenclature

$\epsilon_{\chi\chi}$	strain tensor ($\chi\chi$ component)
ρ	Pearson's correlation coefficient
C_x	covariance matrix
CoV	coefficienf of variation
s_E	experimental shape descriptors
s_M	model shape descriptors
$U(a, b)$	uniform distribution ranging between a and b
$u_(s_E)$	experimental uncertainty
u_{meas}	measurement uncertainty
u_y	displacements y-direction

Chapter 1

Introduction

1.1 Introduction

Computational models are ubiquitous in engineering and science. Depending on their area of application they perform different roles and thus can be divided into two general classes: informative and predictive [1]. The former allow scientists to interrogate physical phenomena that were previously outside their capabilities and can result in the creation of knowledge. The latter are commonly used in engineering, climatology and finance to inform decisions with sometimes substantial socio-economic impacts. It is therefore important that the capability of a model to represent the real world is demonstrated. Towards that goal, Schruben [2] introduced the term credibility which he defined as the 'willingness of persons to base decisions on information obtained from the model'. However, the process of establishing credibility in the results of a model is not straightforward. Oberkampf and Roy [3] suggest that the fundamental elements that build credibility in computational results are: (a) quality of the analysts conducting the work, (b) quality of the physics modeling, (c) verification and validation activities, and (d) uncertainty quantification and sensitivity analyses. This research focuses on utilizing full-field measurements and model predictions to develop metrics that better inform decision makers regarding the validity of their models.

Two important and widely used terms emerged in the previous paragraph while describing the process of establishing credibility in the results of a model:

verification and validation (V&V). Various definitions have been suggested across scientific disciplines, but the ones proposed by the American Society of Mechanical Engineers [4] used by the computational solid mechanics community will be adopted here. In this context, **verification** is the process of determining that a computational model accurately represents the underlying mathematical model and its solution whilst **validation** is the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model.

Verification establishes that the underlying mathematical model of a process is correctly implemented in its computational embodiment. This is achieved by comparing computational solutions against accurate analytical solutions known as verification benchmarks. Verification can be considered the process of answering to the question: are the equations solved correctly? The process of verification precedes that of validation and even though it comprises an important part of establishing credibility in the outcomes of a simulation, it will be assumed that it has been achieved to a certain degree by the companies providing the commercial, modeling software.

Validation on the other hand deals with quantifying the proximity of a model's predictions to the real world and is established through the comparison of model predictions with experimental measurements. In essence, it responds to the question: are the right equations solved? [5] and is the focus of research.

The aim of validation is to assess the predictive capability of the model subject to certain physics assumptions, given a set of criteria or accuracy requirements against which this assessment is evaluated. It is often the case however that following a validation procedure the degree of similarity between the predicted and measured outcomes does not comply with the pre-defined accuracy requirements. In these cases, it is common to implement a calibration procedure to improve the capacity of the model to represent the real world. The process of model calibration which is also known as model updating has been historically associated with the structural dynamics community [6] and can be distinguished into two general domains; a) parameter calibration b) model form revision [4]

Parameter calibration is inevitable in cases where parameters that may or may

not have physical meaning and are used in model building cannot be independently measured (for example to determine the stiffness and damping of mechanical joints). In these cases, calibration may precede validation and the outcomes of this procedure may be subsequently used in larger-scale models. In terms of model form revision, the physics of the model, the associated assumptions and the required level of detail should be well-formulated prior to the building of the model. Inconsistencies between the measurements and the prediction may arise during the validation procedure when it becomes apparent that the embodied assumptions fail in representing the real world even after parameter calibration. In this case, revisions in model form may be considered critical given the existing accuracy requirements. It should be stated that in both cases a new validation procedure should be planned to assess the predictive capability of the model. This practically means the acquisition of measurements and by extension a new experimental campaign within the intended domain where the model has not been previously evaluated. This will ensure that the calibrated model demonstrates the pre-defined capabilities while not being overfit to data.

The aim of model calibration, also known as model updating, is to adjust the model parameters to achieve the best agreement with the measurements and usually succeeds validation. This process should not be conflated with validation where the aim is to assess the capability of the model to represent the real world.

To determine whether the predictions accurately represent the real world, some form of measurement is required. This is achieved by modern measuring equipment which allows scientists and engineers to examine physical phenomena and behaviours with unprecedented level of detail. Compared to traditional analyses where point measurements are used for model assessment, full-field measurements will be employed for the most part of the thesis. The term full-field measurement implies measurements captured across a field or a space, whose post-processing results in a series of values defined over a grid.

Using information-rich spatial data can complicate data handling and analysis. Instead of working with a single measurement or a series of measurements, the practitioner is often faced with large matrices of measurements that may be defined over a grid with density and orientation that may vary across measure-

ments. This abundance in information can discourage practitioners that usually fail to exploit these new capabilities and resort to traditional methods where the quality of a model is assessed on a small sub-region of the overall field via a one-to-one comparison with the model output. One powerful way to address these issues is to employ decomposition techniques. They have been traditionally used to extract features from measurements and predictions, or to compress large datasets. In model validation they have been used to reduce the dimensionality of spatial measurements and predictions, potentially consisting of thousands or millions of pixels by representing them as points in a lower-dimensionality space. These points, widely known as feature vectors, that result from this process allow measurements and predictions to be efficiently compared without any loss of important information.

However, no measurement is exact and the uncertainty in measurements can influence decisions about the reliability of simulations. One of the challenges identified and addressed in this study is the representation of measurement uncertainty, in the low-dimensional form used to extract features from information-rich data fields. This solution offers an improvement to the validation process as it allows model predictions to be quantitatively compared to measurements while accounting for the uncertainty in the latter. The statistical representation of a measurement in its feature vector form can be used for the identification of critical events during temporally evolving phenomena, or to provide information about discrepancies in certain features, thus leading to an improved understanding of the physical process and the limitations of the computational model.

Even though there are many validation metrics that can be used to compare predictions against measurements it is challenging to specify one that can quantify the discrepancy between the two, given the measurement uncertainty in the latter, while managing to sufficiently and simply communicate all the relevant information to non-experts. Various methods have been suggested in the literature that utilize measurements across spatial domains aiming to provide a solution to that challenge ([7], [8],[9], [10], [11]). However, the need for a method capable of accurately establishing a model's accuracy while accounting for the uncertainty in the measurements is still lacking. This challenge has been addressed with the

development of two novel metrics.

1.2 Aim & objectives

The aim of this research is the development of a method to quantitatively assess the quality of a model's predictions while accounting for the associated uncertainties. Compared to traditional methods of model validation where sparse, point measurements are used to evaluate its quality, spatial measurements capable of capturing the real-world response of the modeled physical process across a wide region will be employed. To achieve this aim a series of objectives have been proposed:

- To review the existing validation metrics and determine their potential use with spatial measurements.
- To develop a method that accurately represents the uncertainty-infected spatial measurements into the feature vector space.
- To demonstrate the efficacy and comprehensiveness of the developed techniques on a test case.

1.3 Thesis outline

Following the introduction, a review of the model validation literature is given in **Chapter 2**. This will aid readers to identify the knowledge gaps associated with model validation using full-field measurements. Subsequently, in **Chapter 3** a review of some of the metrics for univariate and multivariate probabilistic model validation is given; their capacities and limitations are identified and the potential use of some of them for model validation using spatial data will be described. In **Chapter 4** a method to transform a full-field measurement and its uncertainty in its feature vector form is proposed. This is accompanied by various applications demonstrating potential uses across scientific disciplines. The technique developed in this chapter is then employed in **Chapter 5** to assess the differences between measured and predicted fields using the Mahalanobis distance.

This comprises the first of two proposed techniques that employ decomposition methods to assess the quality of a model. Examples from the area of structural mechanics will be demonstrated and a comparison with existing methods will take place. Finally, in **Chapter 6** the novelties and capabilities of the various developments will be discussed before reaching **Chapter 7** where the conclusions from this work will be drawn and suggestions for future research outlined.

Chapter 2

Literature review

The aim of this review is to provide an overview of the developments in the area of model validation performed with the aid of field measurements and to identify gaps associated with the existing guidelines and practices. To accomplish that, it will be separated into two parts. In the first part, the topic of model validation will be addressed from a philosophical standpoint and the existing guides for model validation in computational solid mechanics will be reviewed while an overview of the available procedures to characterise the various forms of uncertainty will be given. In the second part, some of the existing comparison operators used to quantify the level of agreement between the predicted and the measured quantities, known as validation metrics, will be analysed along with their assumptions and limitations. This section will be delivered in a manner of increasing complexity, starting from statistical hypothesis testing in univariate problems to comparison methods employed in geostatistics and metrics based on feature extraction techniques for the validation of models using field measurements. The capabilities and limitations of the various techniques will be identified and solutions will be provided in the next chapters.

2.1 Model validation procedures

2.1.1 Validation in the philosophy of science

The problem of model validation has been the focus of attention in the philosophy of science long before modern computational resources were available. The question at hand is whether someone can establish trust or belief in the predictions of a theory. In response, various philosophical currents have surfaced: from the logical positivists of the Vienna Circle [12] who supported the concept that the creation of new knowledge should be based on its empirical verification, to falsificationism supported by Karl Popper [13] in which no theory can be 'verified' but only falsified, i.e. refuted. The latter, in the form adopted by Neyman and Pearson [14] can be regarded as the basis of statistical Hypothesis Testing, where the null hypothesis representing the current status is tested against the alternative hypothesis, given the evidence for a pre-defined level of significance, i.e., the probability of a test rejecting the null hypothesis, given that it is true. On the other hand, Bayesianism [15] considers validation as an empirical process in which a series of successes in model predictions increases the level of confidence in a model or an established paradigm. In the review paper by Kleindorfer et al. [12], the authors demonstrate that model validation is philosophically regressing between two extremes: objectivism and relativism. Objectivism supports the idea that there is a unique basis in which a model or a theory must be resolvable, while in relativism the validity of the model is established relative to established frameworks or standards. They conclude that the validation problem should be converted into an ethical one, in which the practitioner must responsibly and professionally argue for the warrant of the model. In the same way Rudner [16] suggests that the decision to accept or reject a hypothesis depends upon the strength of the evidence which itself is a function of the importance of the output.

Taleb points out [17] that evidence is clearly a probabilistic notion, while Audi [18] supports the concept that an increasing body of evidence tends to increase the degree of belief in the model, following the Bayesian route. It becomes apparent, for the case of physics-based models, that the final decision, to reject or to accept a model's prediction is a function of the amount and quality of evidence and the

significance of making a mistake. This should be taken into consideration during the development of validation metrics, where the various uncertainties should be accurately characterised and the output of the comparison, along with the extent of its uncertainty, should be effectively communicated to decision makers.

2.1.2 Validation guidelines

An early attempt to define a common terminology and methodology for the verification and validation of computational models in engineering was established in 1998 through the AIAA[1] guide for the Verification and Validation of Computational Fluid Dynamics simulations [19]. The terminology was largely based on the earlier DoD[2] 5000.61 Instruction that prescribed procedures for the verification, validation & accreditation of the various DoD models [20]. Afterwards, the computational solid mechanics community via the ASME PTC[3] 60 committee produced the first complete guide [4] clarifying concepts and methodologies that should be in place during the execution of verification and validation activities.

One important aspect of these guides is that they explicitly state the various building processes that should be in place in order to perform a validation activity. Following the ASME flowchart in figure 2.1, two parallel paths can be easily identified; the left which is associated with modeling activities and the right that focuses on experimental testing. It is important to stress that the two paths are independent of each other except at the point of 'Preliminary Calculations' where experimentalists and modelers communicate information such as the type and quality of measurements needed, the boundary conditions and loading application method along with other assumptions that should be taken into consideration during modeling or testing. Another point apparent in this flowchart is that the various uncertainties, both in modeling and experimenting, should be quantified and taken into consideration during validation. These uncertainties, which will be described later in detail, may have an immense effect on the outcome of the validation and should be considered during the decision making process. Finally, having characterised the uncertainties in the simulation and in the experiment,

[1] American Institute of Aeronautics and Astronautics
[2] Department of Defense
[3] American Society of Mechanical Engineers: Performance Test Codes

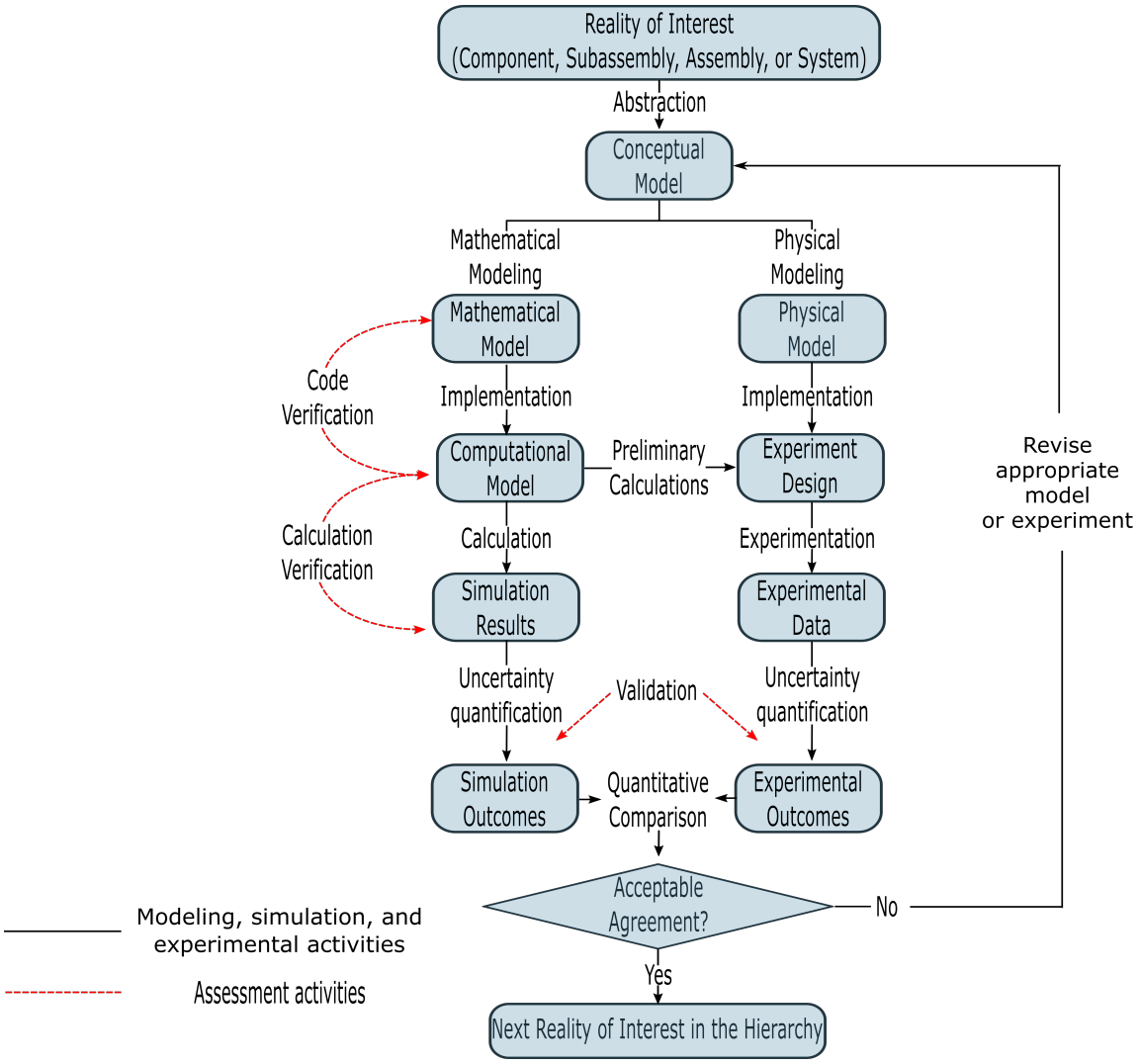

Figure 2.1: ASME validation flowchart [4]

a quantitative comparison of these two datasets takes place. This is performed with the aid of a comparison operator known as a validation metric whose output is compared with pre-defined accuracy requirements. These requirements reflect the level of model-experiment proximity that is defined as being acceptable and should be determined before the commencement of validation activities. The area surrounding the validation metrics has attracted a lot of attention recently and an extensive review will be given in the next section of the literature review.

Although the aforementioned guides provide a conceptual framework across different disciplines by listing a series of suggestions that a practitioner could follow to allow them to establish credibility in their models, a practical step-by-step approach on how to achieve that was still missing. This gap has been partly filled in the form of a guide by the European Committee for Standardization [8].

This guide lists the steps needed by practitioners to characterise the validity of their computational solid mechanics models. This is established by obtaining field measurements of the object of interest using optical techniques such as digital image correlation (DIC), digital speckle pattern interferometry and thermoelastic stress analysis which are later compared to simulated results. Contrary to traditional point measuring devices such as strain gauges or accelerometers that provide a restrictive view of the objects' response, full-field measuring equipment allows the acquisition of a continuous displacement or deformation field across the surface of the object. Another innovation of the CEN guide is that the comparison between the predicted and measured datasets is performed via their feature vector representations. This is achieved via the decomposition of the respective datasets using a set of polynomials like Chebyshev or Krawtchouk, whose elements describe certain features of the data and are known as shape descriptors (SDs). The values of the respective shape descriptors are then assembled in a vector known as a feature vector. The feature vector of the measured dataset is compared to the corresponding predicted one, while accounting for the measurement uncertainty. The latter is obtained during the calibration of the optical system. This comparison is portrayed in equation (2.1).

$$\{\mathbf{s}_M\} = \{\mathbf{s}_E\} \pm 1.96u_E \tag{2.1}$$

$\{\mathbf{s}_M\}$ and $\{\mathbf{s}_E\}$ are the vectors corresponding to the shape descriptors that represent the displacement or strain fields from the model and experiment respectively; and u_E is the experimental uncertainty. The outcome of the validation process is a Boolean acceptance or rejection statement regarding the capability of the model to represent the real-world given the uncertainty in the measurements.

An example of this process is given in figure 2.2 where the two datasets have been decomposed with the aid of Chebyshev polynomials and then plotted against each other. The dashed line corresponds to the ideal scenario where $\{\mathbf{s}_M\} = \{\mathbf{s}_E\}$ while the two continuous lines on either side represent the expanded form of the measurement uncertainty $(1.96u_E) = 0.0196mm)$. Given that all of the shape descriptor combinations are inside the band defined by equation (2.1) the CEN guide suggests that this model prediction is a good representation of the reality of

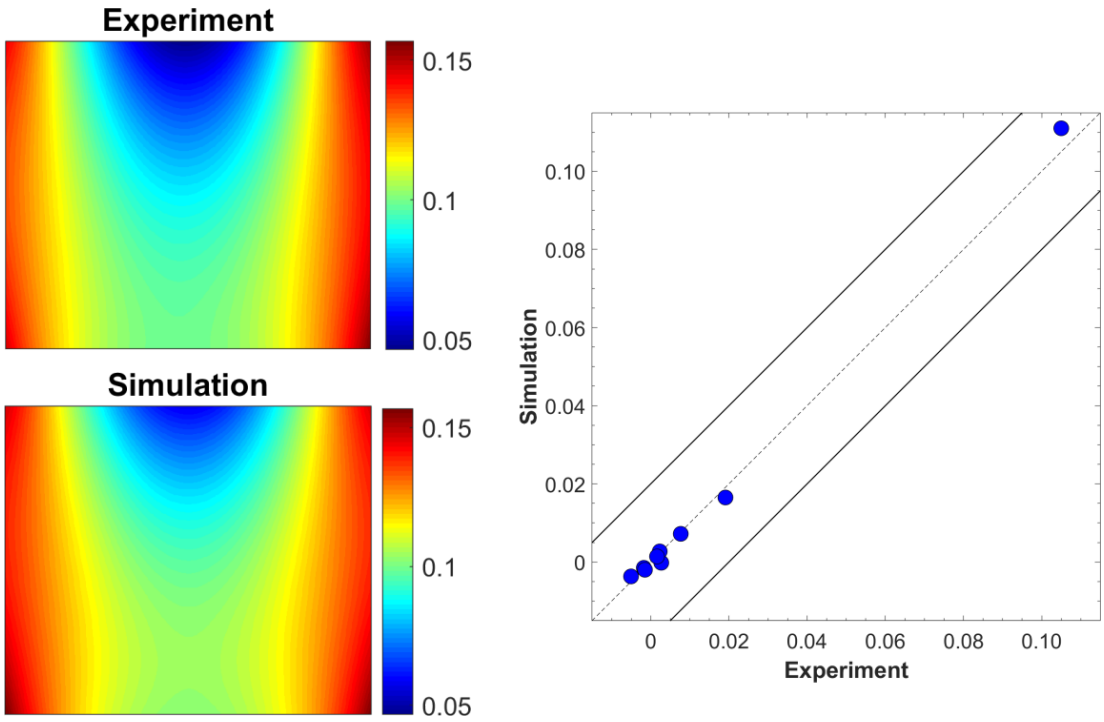

Figure 2.2: CEN-based model validation. The data shown on the left have been decomposed into the respective shape descriptors depicted in figure 2.3.

the experiment. In this example it is qualitatively obvious that the two datasets are similar and the CEN criterion accurately reflects that. The corresponding measured and simulated shape descriptors have also been plotted in the form of bar charts in figure 2.3. The x-axis outlines the shape descriptor identification, a demonstration of which is given in figure 3.9, while the y-axis shows their values. The experimentally acquired shape descriptors have been sorted in descending order of magnitude as reflected by the length of the bars.

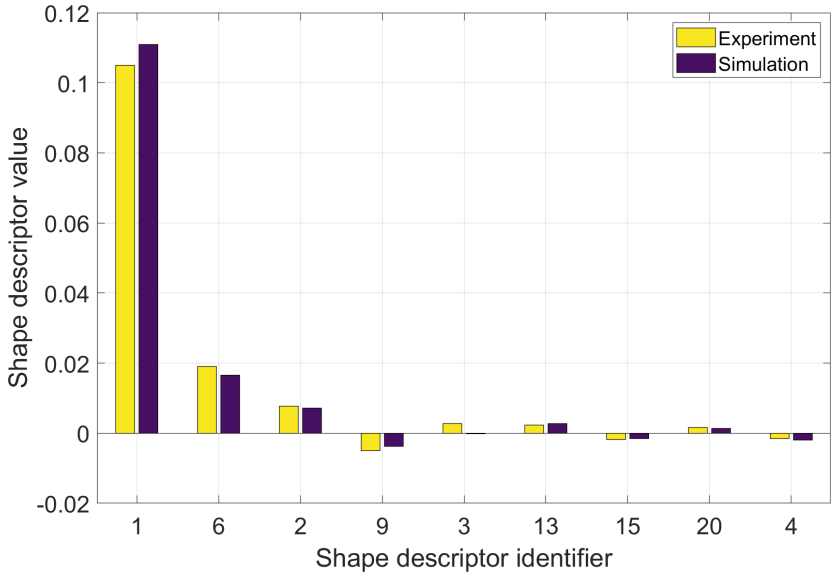

Figure 2.3: Shape descriptor representation of the data shown in figure 2.2.

However, the integrity of this validation outcome is diminished when a different example is considered. In this case, the shape descriptor values of the feature vector corresponding to the experimental measurement depicted in figure 2.2 are perturbed with the aid of a random number generator. The resulting dataset, which now represents the simulation, is shown at the top of figure 2.4 while the CEN-based assessment is demonstrated in the right side of the same figure. Even though all of the shape descriptors are within the band defined by equation (2.1), thus considering the simulation to be representative of the experiment, it can be seen on the left side of figure 2.4 that the magntitude of the pixel-wise differences between the two-datasets exceeds the measurement uncertainty ($u_E = 0.01mm$) by more than five times across extended regions. It is obvious that the magnitude of local differences between the two cannot be attributed to the presence of measurement uncertainty alone. This drawback of the CEN guide should be accounted for.

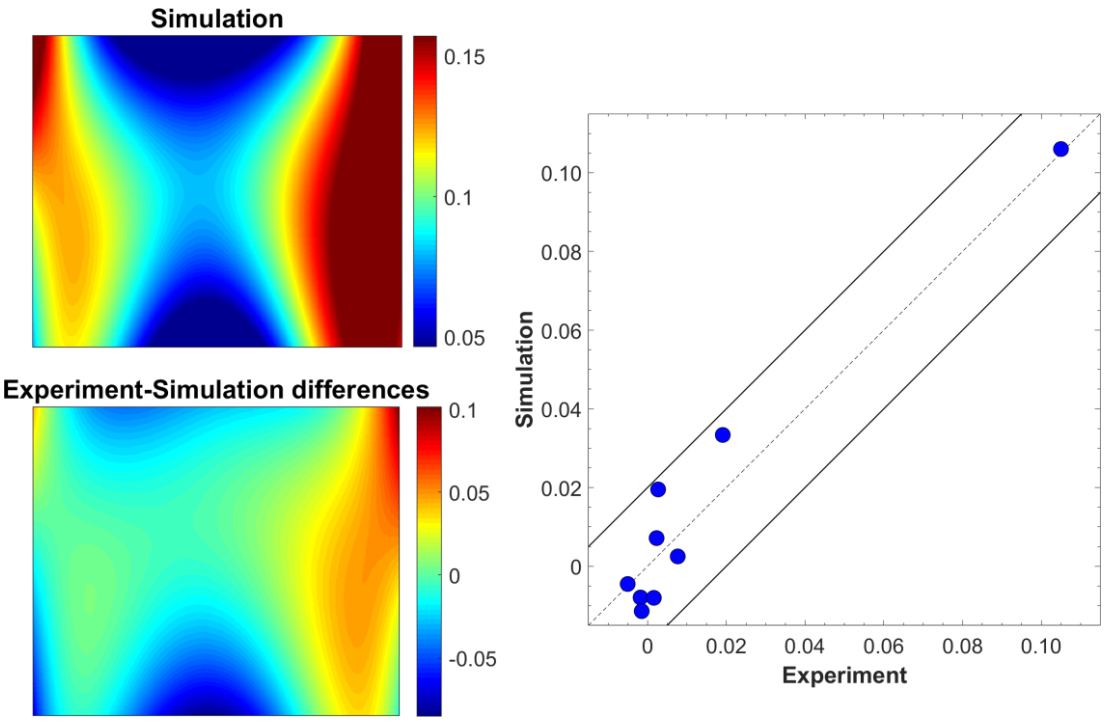

Figure 2.4: CEN-based model validation. The measured dataset is shown in figure 2.2. The simulated dataset is the result of a random perturbation of the measurement's feature vector.

Moreover, this type of assessment (simulation being representative of the real world or not) does not inform decision makers about the capacity of the model to represent reality in a quantitative manner, which is the goal of validation as defined by the ASME guide [4]. Another drawback is that the process can be greatly affected by the quality of the measurements, so that for cases where the measurement uncertainty is negligible, a 'good' model may be rejected, which is known as a Type I error in hypothesis testing, while in cases where the measurement uncertainty is large a 'bad' model may be accepted, known as a Type II error. This could have catastrophic consequences as decision makers will have failed to reject a hypothesis that the model is a true representation of the world when it is not, because of the excessive measurement uncertainty. To establish that this issue does not undermine the validation process, it is stated in the CEN guide that the definition of the measurement uncertainty is a strategic decision for each organisation as its impact is twofold: a) it determines the value of conservatism in the validation (wider bands in figure 2.2 imply more flexibility on what is considered acceptable) b) it determines the cost of the experimentation, as the need for measuring equipment with higher precision can greatly increase

the overall cost.

From these two examples it becomes apparent that there are two missing links from the CEN guide associated with accurate and quantitative model validation. The first, is a method to accurately characterise the measurement uncertainty in the feature vector space, and the second is a method to quantitatively characterise the ability of a model to represent the world.

2.1.3 Uncertainty quantification

The need to characterise the different sources of uncertainty during modeling and experimental testing is pointed out in both the AIAA [19] and the ASME guides [4]. Uncertainties are present in testing, where different measuring devices may lead to different levels of accuracy, in manufacturing, where different processes can lead to substantial differences in geometric tolerances (and cost) and in computational modeling, where changes in the spatial or temporal discretization can severely impact the accuracy of the simulation. Uncertain knowledge during modeling and testing must be identified and properly characterised if decisions are to be made. Failure to do so or selecting an inappropriate method to represent that uncertainty can lead to catastrophic failures [21].

An established dichotomy across the various sources of uncertainty is aleatory and epistemic uncertainty. Aleatory uncertainty is a term commonly used in the context of uncertainty characterization and its use can be viewed from two perspectives, namely practical and epistemological. Practically, aleatory uncertainty is used to characterise the stochasticity or variability in a parameter or process that is considered to be non-reducible, for example when characterizing the stiffness across specimens of a batch. The term non-reducible implies that for the sake of the analysis and due to budget and time constraints (manufacturing processes can be improved but doing so would be expensive) such a reduction is not viable. Traditionally aleatory uncertainty has been characterized using probability distributions. Epistemologically, such a definition could be considered flawed as it implies that that there is no method that can reduce this form of uncertainty and thus cannot be subject to falsification deeming it meaningless [22]. Throughout the thesis the term aleatory uncertainty is used to determine uncertainties that

cannot be reduced within the time frame / scope of the analysis.

On the other hand, the term epistemic uncertainty is attributed to uncertainty arising due to a lack of knowledge and can take various forms, such as ignorance about the type of the distribution that should be used to characterise some phenomenon, or subjective belief about a parameter or quantity. Contrary to aleatory, epistemic uncertainty can be reduced through the addition of knowledge. This may be due to acquisition of more or better quality measurements or sometimes to improved assumptions and updated formulations. Tools used to characterise epistemic uncertainty include Bayesian methods, intervals [23], probability bounds analysis [24],[25] and random sets [26]. It should be noted that the boundaries between these two types of uncertainty are not sharply defined and the categorization of a parameter within one of the two groups is influenced by the problem at hand. Irrespective of their categorization, sound uncertainty characterization methods should be used to accurately treat ignorance and stochasticity without having to resort to unfounded assumptions [27], [28].

Following their categorization and characterization, uncertainties should be then 'propagated' through the model to estimate of their effect on the quantity of interest. Different methods can be used to propagate them; probabilistic approaches include Monte Carlo simulations, perturbation methods and polynomial chaos expansion, among others. Non probabilistic approaches include interval analysis fuzzy theory or evidence theory.

Bayesian methods also play an important role in UQ. Utilizing Bayes rule, they have been used across numerous cases; for example, in parameter calibration where (subjective) prior knowledge is combined with experimental measurements to obtain better estimates [29] for model parameters. This procedure is crucial for parameters whose values cannot be directly measured or the cost to do so is excessive. However, such methods can lead to erroneous results when the model used to represent the physical process is faulty. Kennedy and O'Hagan [30] suggested a framework that can account for the presence of model error, also known as structural error.

To do so, such bias correction techniques employ Gaussian processes that are 'trained' using experimental measurements across a series of validation locations

while allowing an error term to account for the discrepancies between model predictions and measurements. The idea being that models being numerical abstractions of reality are inherently wrong and such discrepancies should be taken into account when making predictions. The advantage of these approaches is that they can also provide estimates of the model's predictive capability at locations in the design space away from the ones used during training. Their disadvantages include various decisions that must be made during their 'training' including appropriate selection of the covariance function and various assumptions, such as normality across validation locations. Although, these techniques seem appropriate for estimating model error, their use lies outside the scope of this work.

Bayesian methods are becoming increasingly popular in the field of uncertainty quantification. This growth has greatly benefited from improvements in computational capabilities that enable computationally intensive techniques such as Markov chain Monte Carlo to be used in a variety of problems. Among such techniques lies approximate Bayesian computation (ABC), a technique that will be employed subsequently to represent the measurement error in the feature vector space. ABC is a relatively new technique developed to allow a posterior distribution to be estimated without knowledge of the likelihood function [31]. When the likelihood function cannot be directly formulated, as in the case of complex structural models, ABC provides an alternative that can be used to represent uncertainty-infected spatial measurements in a lower-dimensional space.

Both frequentist and Bayesian techniques are widely used in model validation. Frequentist inference is based on the relative frequency/proportion of an event compared to the total number of draws/trials. Classical hypothesis testing techniques, confidence-intervals and p-values fall in this category. Bayesian inference on the other hand utilizes Bayes rule to make inferences. To simplify, Bayesian inference represents an interpretation of probability that can be either objective or subjective. Prior knowledge or belief about a process or event (usually in the form of a parameter) is reflected in the prior distribution. Afterwards, the prior distribution is combined with the likelihood (a model that explains the underlying process) and the available data to produce the posterior distribution. This approach, which can be iterative, allows the updating of knowledge about that

process as more data become available. Even though one can argue about their differences from a philosophical standpoint [32] both are used pragmatically to inform decisions surrounding model validation.

It should be stated that the list of UQ methods reported in this section is not exhaustive. Even though UQ is closely related to model validation, a literature review on uncertainty quantification methods lies outside the scope of this work. The interested reader could refer to the works of Smith [33] for a comprehensive treatise on the concepts and methods associated with UQ and to Oberkampf and Roy [3] for a better understanding of the interplay between UQ methods and project planning in the context of industrial-level validation.

Uncertainties in measurements

Procedures to accurately characterise the uncertainty of a measurement are well-established and enshrined in standards, such as ISO 17025 [34] which specifies that the calibration of measuring devices should be achieved through a traceable, continuous chain of comparisons to a primary standard; while in engineering, detailed calibration procedures have been developed, for instance for optical instruments for deformation measurements [35]. The guide to the expression of uncertainty in measurement (GUM) [36] suggests that the uncertainties that can be defined a priori, for instance by calibration, are known as type B; while type A uncertainties are the random component of measurements and can be defined based on a series of measurements or repeated observations. Moreover it suggests that both types of uncertainty can be modelled using probability distributions. In this framework type A uncertainties can be considered as the aleatory form of measurement uncertainty while type B uncertainties can be considered as the epistemic form of measurement uncertainty. Two caveats can be identified in these suggestions. Firstly, the consideration that both uncertainties can be characterised using probability distributions and secondly the lack of guidance when the number of measurements is small (e.g. $\nu = 1$).

Stating that both forms of uncertainty Type A (aleatory) and Type B (epistemic) can be represented using probability distributions is a big simplification, especially for cases where such assumption is unfounded. For example, there

are suggestions within the guide outlining the use of the uniform distribution to represent Type B uncertainties when a specific distribution is not given from the manufacturer of the measuring equipment. Conflating what is known and representing it using distributions that incorporate assumptions that cannot be reasonably justified is wrong; different mathematical forms should be used to represent that knowledge [27].

When the measurement errors are small and random, they can be represented using the Gaussian or Normal distribution [37], which is symmetric and has finite high-order moments. A systematic error or bias in the measurement, represents a constant offset in the measured quantity relative to the true value and in the case of a Gaussian distribution is associated with its mean μ; while the randomness in the measured quantity is related to the standard deviation σ of a Gaussian distribution. This means that when the two values, the mean (μ) and the standard deviation (σ) of the Gaussian distribution are known, as the result of a device calibration process, then one can estimate the true value of a measurement with a certain level of confidence. In simple terms, the mean, μ, defines the peak value in the bell-shaped curve of the Gaussian distribution, while the random error or standard deviation, σ, characterises the width of the curve and to support decision-making a 95% confidence interval can be defined as $[measured\ value \pm 1.96\sigma]$. In those cases where no knowledge about the probabilistic form of the uncertainty exists, the error can be represented using interval analysis [21], in which the confidence interval is replaced with a range that represents the associated uncertainty, i.e. $[measured\ value \pm 1.96u_{meas}]$ where u_{meas} represents the measurement error, usually obtained from a calibration.

2.2 Validation metrics

Having characterised the various uncertainties in the results of the experimental and modeling branches, the next step is to quantitatively compare them, as shown in figure 2.1. This can be performed with the aid of a metric, which during the validation activity is known as a **validation metric**. Oberkampf and Roy [3] define a validation metric operator as a difference operator between computational and

experimental results for the same response quantity. A series of recommendations regarding the formulation of a new validation metric have been provided by Oberkampf and his collaborators ([38], [39]) amongst which are:

- The metric should contain an estimate of the numerical error in the response quantity of interest as a result of the computational simulation.
- A metric should incorporate an estimate of the measurement error in the experimental data.
- A validation metric should be a true metric in the mathematical sense, i.e.,

  Non-negativity : $d(x, y) \geq 0$,

  Symmetry : $d(x, y) = d(y, x)$,

  Triangle inequality : $d(x, y) + d(y, z) \geq d(x, z)$,

  Identity of indiscernibles : $d(x, y) = 0$ if and only if $x = y$.
- The metric should be able to account for cases where there is no uncertainty associated with the measurements and for the cases where quantities may have both aleatory and epistemic uncertainty.

At the same time, it is stated that the metric should be intuitively understandable to engineers, project managers, and decision makers [3]. The significance of this point arises in cases where the final decision, i.e. to reject or to accept a model's predictions for some intended use may be made by a person or a group who may not be involved in the validation procedure or may be lacking the relevant expertise. This characteristic of a validation metric can be the ultimate constraint during the development or endorsement of a metric as its outcome should be widely understandable with no allowances for ambiguity in its interpretation. At this point, it should be stressed that the terms 'metric' and 'distance' are used interchangeably to refer to the notion of a validation metric as a difference operator without necessarily adhering to its strict mathematical sense.

2.2.1 Univariate case

A number of different metrics have been developed that can be used to characterise the validity of a model given a response quantity, e.g. displacement or acceleration at a point, or characteristic, e.g. mean displacement across a field. In the review paper by Liu et al. [40], the most popular metrics for stochastic model validation are compared, i.e. for cases where the simulation and experimental outcomes take the form of probability distributions. The reviewed approaches include: hypothesis testing, Bayesian factor, frequentist's metric, area metric. Even though these approaches can be extended to the multivariate space, here only their univariate forms will be addressed for simplicity.

Hypothesis Testing [41] has been traditionally employed to test for statistically significant differences between two parameters (e.g. mean or standard deviation with the aid of the z and F-test respectively) or distributions (with the aid of the Kolmogorov-Smirnov [42] [43] or the Anderson-Darling tests [44]). In model validation, hypothesis testing is associated with the comparison of a simulation (usually in the form of Monte Carlo outputs) to a series of empirical observations. Hypothesis testing consists of two competing hypotheses; the null hypothesis (H_o) that represents similarity between the measurements and predictions and, the alternative hypothesis (H_a) that represents a significant difference between the two. The output of the comparison is deemed statistically significant if the alternative hypothesis is unlikely to realize due to randomness alone for a pre-defined threshold (the level of statistical significance). In the case that there is not enough evidence to support the alternative hypothesis, the outcome can be quite ambivalent as it reflects a failure to reject the null hypothesis, which is different from confirming it; this output poses a severe limitation when used in model validation where the objective is a quantitative assessment of the extend of deviation between measurements and simulations.

Another assumption often incorporated during hypothesis testing is that the measurements and predictions are normally distributed. When this is not the case (non-Gaussian, skewed distributions), appropriate transformations such as the Box-Cox power transformations could be implemented [45]. It should be noted, however, that such transformations have certain limitations (e.g. the ran-

dom variable should be positive) and they do not ensure the normality of the transformed variables. Moreover, for the cases where a small number of predictions or measurements is available, the power of the various tests is hampered thus resulting in Type II errors [44].

The **Bayes factor** can be viewed equivalent to hypothesis testing implemented in a Bayesian setting. Bayesian analysis is based on the Bayes rule, which for the case of a parameter θ given some measurements/data D, is given by equation (4.1)

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta} \tag{2.2}$$

where $p(\theta|D)$ is the posterior distribution for θ given the data D, $p(D|\theta)$ is the likelihood and $p(\theta)$ is the prior distribution, while in the denominator $\int p(D|\theta)p(\theta)d\theta$ is a normalizing factor known as the marginal likelihood. Bayesian inference is the reallocation of credibility across possibilities [46]. This is reflected in the posterior distribution via updating the prior distribution through the likelihood. The prior distribution represents the existing knowledge about the parameter θ prior to the acquisition of the empirical observations, while the likelihood represents the fitness of these observations to the model for the parameter value θ.

The Bayes factor is then defined as the ratio of two likelihoods: the first likelihood, $p(D|H_1)$ that represents the null hypothesis and the second, $p(D|H_2)$ that represents the alternative hypothesis and its mathematical form is given by equation (2.3).

$$B_{12} = \frac{p(D|H_1)}{p(D|H_2)} \tag{2.3}$$

Rebba and Mahadevan [47],[48] have implemented Bayes factor in a model validation setting for the comparison of point estimates or whole distributions. The first is based on establishing intervals representing equivalence between the estimates (such as means or standard deviations) while the latter is based on comparing the null hypothesis that the prediction is true against the alternative that it is not. Its advantages are that it can produce a quantitative value representing the ratio of 'evidence' of one hypothesis against the other and it

can be used even for cases where a small number of measurements is available. Its disadvantage is that its value can only be subjectively interpreted. Different thresholds regarding the strength of the evidence of one hypothesis against the other have been suggested in the literature by Jeffreys [49] and Kass and Raftery [50], thus making the decision to select one hypothesis over the other quite fuzzy.

The **Frequentist's metric** [51] is a special case of classical hypothesis testing where the difference between the model and experiment sample means across multiple validation sites is established using confidence intervals. The term validation site is used to describe the setting/environment in which the validation process is performed. This can be characterised by the loading and boundary conditions at a specific point in the test procedure and can be extended to more than one site. The aim of the validation metric is to aggregate the comparisons at every site in a global metric that can be effectively communicated to decision makers. For the case of the frequentist's metric, even though it allows practitioners to establish confidence intervals at multiple validation sites, it only considers the mean as the way to do so, thus nullifying the effect of higher-order moments (e.g. variance, kurtosis, skewness)

It has been shown [40] that the **area metric**, [52] allows for a quantitative model-experiment comparison at a single validation site, via the calculation of the absolute area between the predicted and the measured empirical distribution functions for the response quantity of interest. An empirical distribution function (EDF) can be described as a non-parametric way to summarise measurements or predictions and is similar to the cumulative distribution function. More on the theory underlying empirical distribution functions will be given in the next chapter. The area metric calculation is reflected by the grey area in figure 2.5 where a normal distribution $\mathcal{N}(0,1)$ (simulation), in its cumulative distribution function form, is compared against a series of 'measurements', shown in the form of an empirical distribution function, drawn from the same distribution.

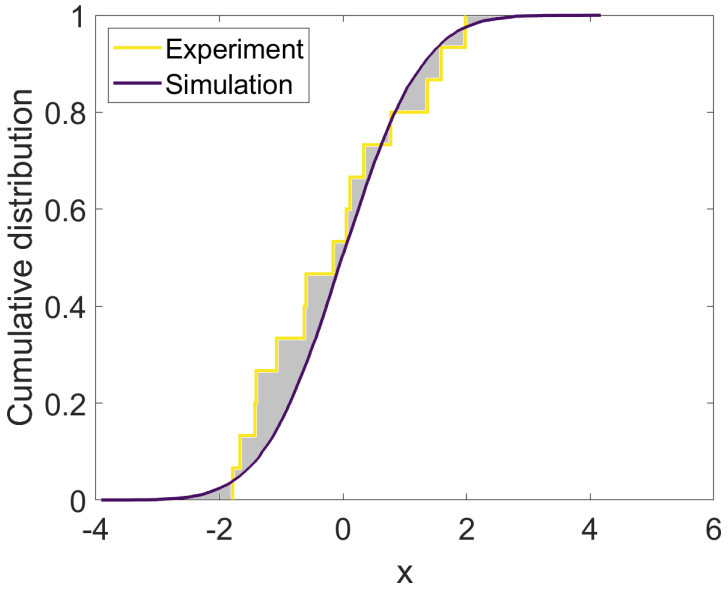

Figure 2.5: Measurements and predictions plotted as empirical distribution functions. The predictions are depicted by the continuous purple line, while the measurements by the yellow stepped line. The grey area between them constitutes the output of the area metric.

Some of the advantages of the area metric are that it can identify differences between models with higher or lower variability while it can also provide a quantitative comparison for cases where a single model output is compared to a series of measurements or where stochastic model output is compared to a single measurement and the associated measurement uncertainty is zero. Moreover, the physical units of the problem are retained thus making it particularly relevant in engineering applications.

Zhan et al. [53] used the area metric to assess the capacity of three surrogate modeling techniques to represent a finite element model in a vehicle design application. In order to assess their models across different validation sites (temporally evolving), they calculated the average area metric. Even though the outcome of this calculation is quantitative it does not reflect the impact of emerging temporal correlations. Bredbenner et al. [54] used the area metric to assess the proximity of finite element models of a set of cervical spines subject to geometric and material variations, compared to experimental measurements. They implemented a normalised version of the area metric that allowed them to assess model-experiment deviations across a wide range of loading levels. They concluded that the area metric can assess the deviation between the measurements and predictions, how-

ever since no established standard of what constitutes an acceptable model exists, the outcome of the comparison should be subjectively assessed on the goals and objectives of the use of the model.

Ferson et al.[52] also proposed a way to integrate predictions and measurements across multiple validation sites via a technique known as **u-pooling**. The idea is that multiple measurements can be aggregated and assessed against the model's predictions using the latter as the base for their transformation. Analytically, the measurements are transformed into u-values using the model's empirical distribution function in the following manner: $u_i = F^m(y_i^e)$. The EDF of the u-values is then plotted against the cumulative distribution function of a uniform distribution defined in the domain $[0, 1]$ and the area metric between the two is calculated as demonstrated in figure 3.4. The use of the cumulative distribution function of a uniform distribution stems from the probability integral transform (PIT) [55] that states that a variable defined through the CDF $F(X)$ of a continuous random variable X is uniformly distributed in the domain $[0, 1]$. The process can be graphically seen in figure 3.4. The value of u-pooling is bounded between 0 and 0.5 with the former demonstrating a perfect agreement between the measured and predicted distributions and the latter meaning that no similarity between them is evident. Unfortunately, the interpretation of this result is subjective as will be demonstrated in the following chapter.

U-pooling has been widely implemented to assess the discrepancy between probabilistic models and measurements. In [56] it was implemented to aggregate measurements corresponding to three different tire tread designs and assess them against model predictions. In addition to this, the authors combined u-pooling with an innovative technique that allowed them to account for the limitation in experimental measurements within hypothesis testing. The outcome of their technique is an assessment of the significance of the validation outcome (reject or not the validity of the model), for a pre-defined significance level. In a similar way Gorguluarslan et al. [57] combined u-pooling with the Kolmogorov-Smirnov hypothesis test to assess the model's accuracy to simulate the mechanical behaviour of lattice structures used in bone implants. To characterise the validity of their models they used the Kolmogorov-Smirnov test to test for equality between the

u-transformed EDF and the CDF of the uniform distribution. Even though this technique relieves the difficulty of interpreting the outcome of u-pooling and it is associated with traditional hypothesis testing, it should be applied with careful consideration, especially when the number of available measurements is small. It was shown by Razali et. al. [44] that the power of the Kolmogorov-Smirnov test (probability of correctly rejecting the null hypothesis when the alternative is true) is diminished for small sample tests (as in this case $n_E = 4$). In these cases, failing to reject the null hypothesis (the two distributions are equal) does not necessarily mean that the model is valid.

Gardner et al. [58] in their review paper highlighted metrics that could be used to quantify discrepancies between probabilistic distributions corresponding to predictions and measurements respectively. Even though the authors defined the criteria that would render the various metrics desirable within an engineering context it is apparent that most of them, including f-divergences such as the Kullback-Leibler and integral probability metrics such as the Kolmogorov-Smirnov distance and the Maximum Mean Discrepancy, are quite hard to objectively interpret and also require a large number of samples to accurately assess discrepancies between distributions. Even though some of them, such as the Maximum Mean Discrepancy, can provide decision makers with additional information, including a visual representation of the discrepancy between distributions, it is not straightforward to apply these metrics in cases of high dimensionality.

2.2.2 Multivariate case

Compared to the previous section, where the validation of a model was limited to a single response quantity, e.g. displacement or strain at a point of interest, there may be cases where the validity of the model should be established through the integration of multiple response quantities, such as frequency and mode shapes [59].

The u-pooling technique for aggregating multivariate stochastic outputs, i.e. accounting for different response quantities at the same time, can only be used when these outputs are independent of each other. Li et al. [60] identified this short-coming and extended its application to account for the cases where correl-

ations between variables may occur. To do so they developed the **Probability Integral Transform (PIT) area metric** which is based on the Multivariate Probability Integral Transform [61] and quantifies the disagreement between two multivariate probability distributions. The procedure followed is the same as in u-pooling, the only difference being that the u-transformation of the measured quantities, is based on the multivariate predicted distribution, and the transformation of the predicted quantities in the probability space is no longer uniform but follows a distinctive pattern. This pattern can for example reflect the amount of correlation between the variables as shown in figure 3.39 for the 2-D case of a Gaussian distribution. After the transformation of both datasets, into the newly formed curves, their differences can be quantified using the area metric [52]. The output of the comparison is now bounded between $[0, 1]$. A detailed explanation of this method along with a graphical demonstration will be given in the next chapter. A drawback of this method, similar to u-pooling, is the difficulty in interpreting its output in a decision-making framework along with the need for a big number of model runs to accurately sample from the predicted distribution as the dimensionality of the problem increases. Moreover, as it will be demonstrated in the next chapter, its capacity to accurately assess the model-experiment agreement is impeded in the scenario where the predicted outputs are negatively correlated or the measurements demonstrate some directional bias with respect to the simulation outcomes.

Among the metrics used for the comparison of multivariate probabilistic forecasts is the **Mahalanobis distance (MD)** [62], [63], [64]. The Mahalanobis distance is an extension of the Euclidean distance in the multivariate, d-dimensional space that measures the distance between a point $\{\mathbf{x}_i\}$ and the sample mean $\overline{\mathbf{x}}$ that comprises the distribution, given its sample covariance mareix, $\mathbf{C}_x$. The Mahalanobis distance, in its general d-dimensional form is given by equation (2.4). The total number of samples comprising the distribution will be considered to be N and the number of dimensions d. The size of the sample covariance matrix will then be $d * d$.

$$MD_i = \sqrt{(\{\mathbf{x}_i\} - \{\overline{\mathbf{x}}\})^T [\mathbf{C}_x]^{-1} (\{\mathbf{x}_i\} - \{\overline{\mathbf{x}}\})^T} \tag{2.4}$$

As the name suggests the covariance matrix stores the variances and covariances of the variables in the multivariate distribution. In its 2-dimensional form it is defined as

$$C_x = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{2.5}$$

where the elements of the main diagonal represent the variances of each variable while the rest, off-diagonal elements, represent the variables' covariances; the latter can be defined as the product of their standard deviations and their correlation coefficient, ρ_{12}. For the bivariate case the MD is given by eq. 2.6

$$MD_i = \sqrt{\left(\frac{x_{i1} - \overline{x_1}}{\sigma_1}\right)^2 + \left[\left\{\left(\frac{x_{i2} - \overline{x_2}}{\sigma_2}\right) - \rho_{12}\left(\frac{x_{i1} - \overline{x_1}}{\sigma_1}\right)\right\}\frac{1}{\sqrt{1-\rho_{12}^2}}\right]^2} \tag{2.6}$$

It is obvious that if the correlation between two variables is 0 ($\rho_{12} = 0$) then the MD is equal to the Euclidean distance from x_i to $\overline{x}$ where each component is normalized by its standard deviation. In that form, MD portrays the distance, in standard deviations of a point x_i away from the mean of the distribution. For the limit case where the variables are independent of each other and their standard deviations are equal to one, the MD becomes the Euclidean distance. A demonstration of the MD is shown in figure 2.6 where the bivariate probabilistic outputs (displacement and strain) of a Monte Carlo simulation are shown as purple-coloured circles. Their scattering is captured in the covariance matrix and the amount of correlation is reflected by the ellipses depicting MD isocurves. It is obvious that the further a point is located away from the mean (here the centroid of the 'cloud' of predictions) the greater its Mahalanobis distance.

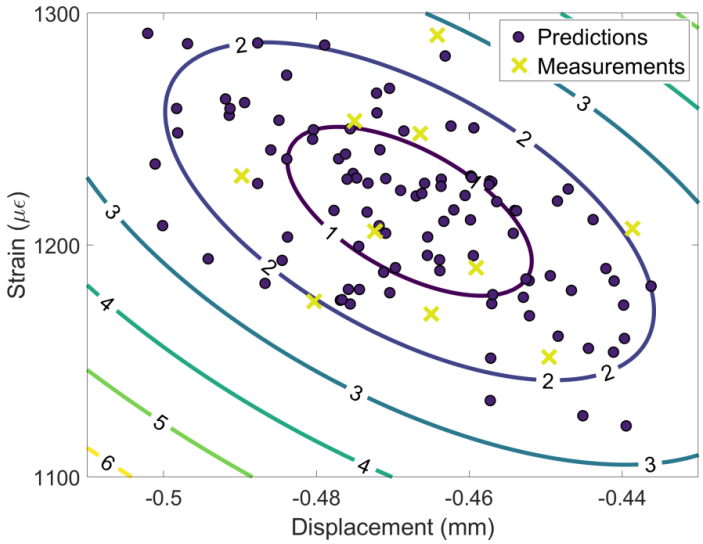

Figure 2.6: Visualisation of the Mahalanobis distance in 2-D space.

The MD has been used extensively in the model validation literature. Bi et al. [65] used it to calibrate and subsequently validate a finite element (FE) model of a steel structure using the first four natural frequencies of the structure as features. They also proposed the use of the pooled MD that extends the capability of the MD in that it allows the characterization of the distance between two distributions, i.e. the measured against the predicted. This is calculated by modifying the covariance matrix so that its values are a weighted mixture of the respective matrices and the weights are defined by the number of samples in each distribution. Zhao et al. [66] used the MD to transform multivariate model predictions into a univariate empirical distribution function by calculating the MD for each of the distribution points and then used the area metric between this and the similarly MD transformed measurements to quantify their agreement. This simplifies the analysis as the dimensionality of the problem decreases to 1-D and is demonstrated in figure 3.8. Their technique will be explored in more detail in the next chapter. In a similar manner, they extended the u-pooling technique for multiple validation sites thus allowing them to account for the correlations between variables. Hu et al. [67] used the same technique but, instead of using the area metric to quantify the agreement between the two resulting distributions implemented the Bhattacharyya distance [68]. The Bhattacharyya distance is a

measure of similarity between two distributions and is defined as

$$d_B(p,q) = -\ln\left(BC(p,q)\right) \tag{2.7}$$

where p and q are the probability distributions that are compared and BC is the Bhattacharyya coefficient. The latter is defined as:

$$BC(p,q) = \sum_{x \in X} \sqrt{p(x)q(x)} \tag{2.8}$$

The range of the Bhattacharyya distance is: $0 \leq D_B \leq \infty$ while the range of the Bhattacaryya coefficient is: $O \leq BC \leq 1$. Finally, the Mahalanobis distance, which is used to compute the Hotelling's t-squared statistic (t^2), a multivariate extension of Student's t-statistic, was used by Balci and Sargent in [69] to test for equality of means across two multivariate normal distributions simulating a queuing system.

Other measures, such as the Kullback-Leibler divergence [70] and the Bhattacharyya distance [68] have been used to quantify the mismatch between multivariate distributions with engineering applications in [71] and [65] respectively.

2.2.3 Accuracy requirements during validation

An important component of validation that has been greatly neglected in the scientific literature is the definition of the decision boundary used to characterise the validity of a model. This boundary is known as accuracy adequacy [4] or as accuracy requirements [3] and allows decision makers to identify whether the model is representative of reality for an intended use. Or as Oberkampf and Roy suggest ([3], pp. 30): "Without accuracy requirements, the question: How good is good enough?" cannot be answered.

Accuracy requirements can take various forms. In engineering they are usually shaped as tolerance intervals within which a model is considered to be valid, i.e. predictions should be within 5% or 10% of the measured quantity, or as in the CEN guide for the validation of solid mechanics models, using full-field measurements, can be defined with respect to the uncertainty in the measurements. In hypothesis testing, the accuracy requirement is represented by the selected α

value, known as the level of statistical significance, which expresses the probability of falsely rejecting a hypothesis while it is true. This notion, along with Type I and Type II errors, was introduced by Neyman and Pearson [72],[73] from a decision-oriented perspective contradicting that of Fisher's significance tests, where the outcome of the test was supposed to be used as an indicator of whether a rare event has occurred or to identify whether the hypothesis is wrong, with no consideration of an alternative hypothesis [14].

Various solutions have emerged across different disciplines while addressing the issue of accuracy requirements. In pharmacology, two different drugs or formulations of the same drug are called bioequivalent if they are absorbed into the blood and become available at the drug action site at about the same rate and concentration [74]. The problem is demonstrating to the regulatory agencies that the new drug has similar effects compared to the 'brand'-drug. **Equivalence tests** [75] incorporate the notion of accuracy requirements in the following way: compared to traditional hypothesis testing where the null hypothesis represents the case that there is no difference between means (prediction vs experiment) against the alternative hypothesis that there is a statistically significant difference between the means, the order is reversed by establishing that the null hypothesis is that the model does not meet the accuracy standard, i.e. there is a difference between the two means given a range of equivalence and the alternative is that it does, i.e. there is statistically significant similarity. As Robinson and Froese [75], pp. 905 suggest: "the equivalence test shifts the burden of proof onto the model", in that the provided evidence acts towards the rejection of the null hypothesis, which is that the model is not valid, or as in pharmacology, that the new drug is bioequivalent to the existing brand drug. The notion of equivalence tests, which was initially restricted to the comparison of sample means, was later expanded by Robinson et al. [14] to test for similarity between two series of observations, thus allowing for the comparison of two distributions, i.e. predictions against measurements. This is achieved via a regression-based technique, where the search for statistically significant similarities is established on two tests for equality; one for the intercept and one for the slope of the regression given pre-defined equivalence ranges for both.

Rebba and Mahadevan [76] proposed a similar idea to account for accuracy requirements but from a different point of view. They proposed that a model is accepted only when the probability of the random variable D, which is defined as the difference between predictions and measurements lies within a pre-defined range $(-\epsilon, \epsilon)$, is larger or equal to a constant c. Or, a model prediction is accepted when:

$$P(-\epsilon < D < \epsilon) \geq c \tag{2.9}$$

The authors demonstrated that the suggested metric, named the **reliability metric** can be easily extended to multivariate distributions and its calculation, for comparing discrepancies between means can be calculated either, via resorting to the central limit theorem [41] for cases where more than thirty independent samples are available or via the calculation of bootstrap estimates for sample statistics. Thacker and Paez [77] proposed the **Z-metric**, which is similar to the reliability metric with the only difference being that the D variable representing the difference between predictions and measurements is normalized by the latter. Its advantage is that for given accuracy requirements a simple statement of the form 'There is a ... % probability that the error between the model and test is not be greater than ... %' thus combining a probabilistic statement with percentage differences which can be easily understood by non-experts.

2.2.4 Validation approaches incorporating field measurements

After providing an overview of the available validation metrics for the univariate and multivariate cases, now the focus of the literature review will move to the validation of computational solid mechanics models using field measurements. Even though, modern, full-field measuring equipment such as digital image correlation have greatly influenced the way that model validation is practiced in engineering, and innovative ways to quantify the level of agreement between the model and the experiment have emerged, spatial data have been traditionally used across a number of disciplines, some of which are: meteorology, climatology,

oceanography, hydrology, ecology and geography. It would then be beneficial to identify some of the key model validation techniques that have been used in those domains and could potentially be transferred to engineering.

The geostatistical approach

A commonly cited quote when dealing with spatial data is Tobler's first law of geography: "everything is related to everything else, but near things are more related than distant things" [78], pp. 236. This dictum is used to stress the fact that data in a spatial domain are not independent but autocorrelated. The autocorrelation or 'smoothness', describing the magnitude of spatial correlation, in a dataset implies the existence of redundant information. This means that even if the measuring equipment is capable of providing high spatial density measurements, the amount of new information in each grid location will be relatively small if the underlying dataset is highly correlated, or as Griffith [79] suggests: when the spatial autocorrelation is zero (i.e. spatial data are independent of each other) the effective sample size, n^* is equal to the number of grid points in the dataset, n, i.e. $n^*{=}n$. On the other hand, when the data are perfectly spatially correlated the effective sample size of the dataset is equal to one, i.e. $n^* = 1$. This situation affects the way that sample statistics are calculated and subsequently the output of hypothesis testing.

Different methods have surfaced to tackle this issue, many of which are in the field of geostatistics. In that domain, the spatial dataset is treated as a 'realization' of a random (stochastic) process. This means that the measured dataset is a realization amongst an infinite number of possible realizations, of an underlying process. However, to be able to make use of the developments in that domain, the assumption of stationarity is usually evoked. This means that the dataset is now characterised as a spatial process whose mean is assumed to be constant and independent of the spatial location and the covariance between locations is defined by their pair-wise distance only (second-order stationarity). However, these assumptions may not always be realistic; for example, when there is a trend across the region, the mean of the underlying process cannot be considered constant and de-trending should take place before moving on the analysis. In addition to

this, no formal test of stationarity in data exists and adjusting for cases where stationarity cannot be rightfully assumed may not be obvious [80]. In cases where stationarity is reasonable though, techniques such as Kriging [81] which in broad terms, is a process of finding the best statistical model fit to the given data. allow the modeling of spatial interactions.

It should be stated however that Kriging has a wide range of applications outside geostatistics, where stationarity is not a requirement. Many of those applications use Kriging, also known as Gaussian Process regression (GP), to build data-driven models or to speed up complex simulators by substituting them as emulators. The benefits of Gaussian Processes include uncertainty quantification for predictions, the need for little a priori knowledge during their building and the capacity to exploit Gaussian properties during analysis, deeming them desirable in decision-critical applications. An application along with a detailed explanation of GPs can be found in [82] where a complex high-dimensional epidemiological HIV simulator was successfully emulated with the aid of GP modelling.

Clifford et al. [83] introduced the notion of the effective sample size, while assessing whether the correlation between two spatial variables, is statistically significant. To do so they made a series of assumptions, including process stationarity, and stochastic independence of the parameters X and Y. Dutilleul et al.[84] provided an improvement to the work of Clifford et al.; while more recently, Griffith [79] developed a method to determine the effective sample size during the calculation of the spatial mean for normally distributed data and then provided extensions for bivariate sample means. Li et al. [85] used the techniques developed by Griffith to visualize relations between stochastic variables using scatterplots and resampling techniques that allowed them to retain the statistical characteristics of the datasets without having to plot the high amounts of redundant information that result from spatial autocorrelation.

It should be emphasized that to be able to make use of all the information provided by full-field measuring devices in a decision making framework, like model validation, an accurate characterisation of the uncertainty in the extend of similarity should be in place. Even though there has been a series of developments in geostatistics that can be used to characterise the uncertainty about

that similarity (e.g. correlation coefficient between two datasets) while accurately accounting for the effective degrees of freedom of the datasets analysed, there are some limitations that currently deter their application in model validation. Firstly, a meticulous analysis of the datasets should take place to identify trends and to remove them before modeling their covariance structure. This means identifying some linear or higher-order polynomial to represent the signal in the data while avoiding overfitting. After that, the need to model the covariance structure adds another level of complexity and subjectivity as some type of model validation should be incorporated; this time to test whether the underlying statistical model accurately describes the behaviour of neighbouring regions. To conclude, the number of assumptions and overall complexity of analysing data using geostatistics does not appear to be of benefit for model validation, at least for the time-being. However, it is my belief that the continuous stream of developments in the domain will eventually provide a solution to the problem of characterising the similarity of two spatial datasets with an accurate degree of confidence.

Pixel-wise comparisons

A different approach was suggested by Levine et al. [86] which was based on pixel-wise comparisons across spatial maps in ecology. Analytically, they wanted to identify the importance of various parameters on species distributions using ecological niche models. To do so they employed a series of metrics, including the calculation of the percentage of pixels whose absolute differences were one standard deviation larger than their mean difference. Wilson [87] used distance measures, such as the Euclidean distance, Kullback-Leibler divergence and Hellinger distance, to quantify the pixel-wise differences during the analysis of maps also produced by species distribution models.

More recently, Jones et al. [88] and Wiederholt et al.[89] used the **(s)tructural (sim)ilarity (SSIM)** index to compare spatial maps of species distribution data and ecological restoration scenarios respectively, while Robertson et al. [90] used it as a discrepancy measure to check for model fit in a Bayesian spatial modeling framework. The SSIM index was developed by Wang et al. [91] to characterise

the quality of image compression in a manner that resembles human visual perception. One of the main points in their paper is that traditionally used metrics such as the mean-squared-error (MSE) fail to accurately characterise the difference in quality of compressed images when compared to the reference image, thus resulting in cases where images of different quality (humanly perceived) may have the same MSE. In practice, the SSIM is calculated as the product of three measures: luminance, contrast, and structure, which respectively account for the difference between means, variances and covariances of the compared datasets and are supposedly independent.

This produces three distinct datasets via the application of a local Gaussian weighting function. The values of the final dataset, which is the product of the three, range between -1 and 1, with the former meaning perfect dissimilarity and the latter perfect similarity. The mean values of that dataset can then be used as a metric of similarity. Compared to previously mentioned techniques, this measure can account for image misalignment through the structure measure, a feature that is beneficial during the comparison of datasets that demonstrate high levels of spatial variation (e.g. strain field around a crack) and slight misalignments can lead to wrong judgements.

An example outlining the potential use of the SSIM index for validation with full-field measurements is demonstrated in figure 2.7. The structural similarity between the measured dataset of figure 2.2 and the simulated one of figure 2.4 is shown at the bottom right. The mean of the SSIM is 0.954 suggesting almost excellent similarity between the two datasets. The biggest deviations stem from the luminance which accounts for differences in the means and is defined as:

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{2.10}$$

C_1 is a constant used to avoid numerical instability when the sum $\mu_x^2 + \mu_y^2$ is close to zero. The authors suggest that $C_1 = (K_1 L)^2$ where L is the dynamic range of the pixel values and $K_1 \ll 1$ is a small constant (the authors used the value of 0.01 in one of their examples).

It can be seen that the largest deviations are located at the bottom of the region, while smaller deviations on the top and towards the right side are visible.

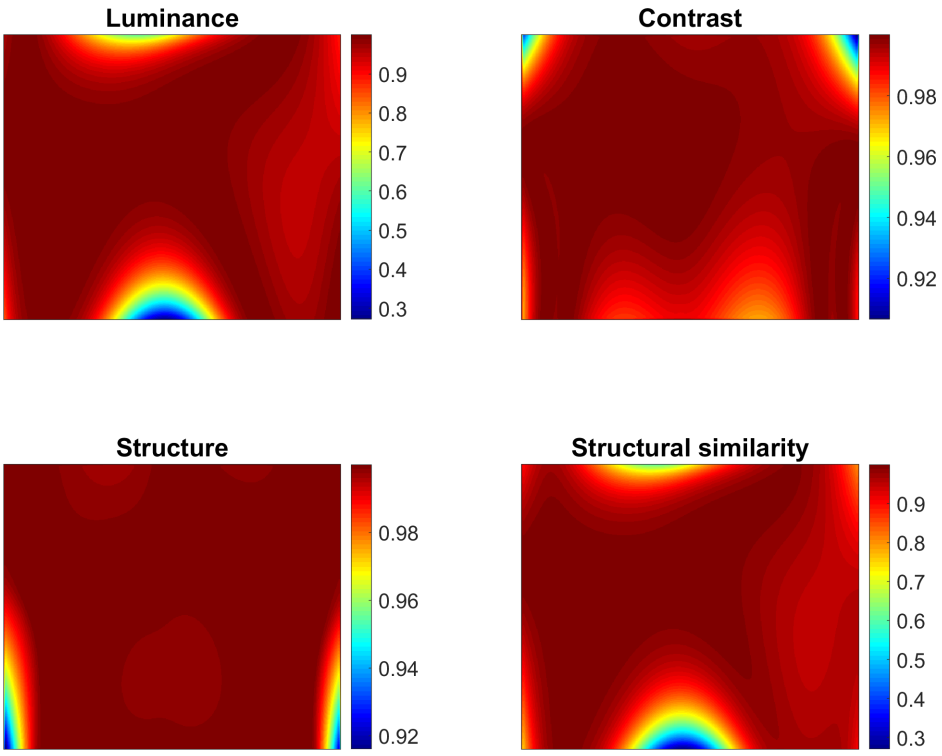

Figure 2.7: Luminance, contrast, structure and their product depicting the structural similarity for the spatial datasets shown in figures 2.2 and 2.4.

The regions demonstrating the largest deviations are in agreement with the pixel-wise differences shown in figure 2.4. The map of contrast describes the differences in the variances of the two datasets and is defined as:

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{2.11}$$

$C_2 = (K_2 L)^2$ and $K_2 \ll 1$ for the same reason as above. It is apparent from the range of the colour bar that the two datasets locally have the same variances, except the two top edges where the values are smaller. This can be attributed to the high dynamic range of values of the simulation in the same regions.

Finally, the structure reflects the correlation between the two datasets and it is defined as:

$$c(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{2.12}$$

$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$ is the covariance between the two signals and $C_3 = C_2/2$.

It can be concluded that the structural similarity can be used to extract valu-

able information from the comparison of two spatial datasets. However, the outcome of the assessment (bounded in $[-1, 1]$) can only be subjectively interpreted. Moreover, further steps should be taken to accurately represent the measurement uncertainty or variability in the dataset. Another issue is the definition of the radius of the Gaussian weighting function. Even though the authors suggest a value of 1.5 (pixels), selecting larger values can greatly impact the outcome of the comparison.

2.2.5 Feature extraction based validation approaches

Even though modern-measuring devices can provide measurements of unprecedented quality and spatial density, it is not obvious how these information-rich datasets can be used by practitioners to assess the validity of their models. A commonly employed technique is to focus on a region of the object and then extract data across a linear segment, which will be later used for validation. This method has been employed, for example, in [92] to validate a voxel-based finite element model of a human mandible using digital speckle pattern interferometry, in [93] to validate a finite element model of the proximal femur using digital image correlation and in [94] for the validation of mesoscale FE analysis of textile composites using digital image correlation. Depending on the area of application, this technique can lead to wrong conclusions with disastrous consequences as only small portions of the overall dataset are taken into consideration, failing to exploit the capabilities of the new technologies.

One powerful way to address these issues is to employ feature extraction techniques that allow the representation of data without any loss of important information. A very popular technique used to produce feature vectors via dimensionality reduction is Karhunen–Loève decomposition, also known as principal component analysis (PCA). PCA is amongst a family of techniques commonly used to describe processes and random vectors; from vibration analysis in [95] to fluid mechanics where Lumley and his co-workers [96],[97] used proper orthogonal decomposition to characterise the coherent structure of turbulence. More recently, weighted proper orthogonal decomposition has been used to generate reduced order models of swirling flow from a turbine [98] and dynamic mode de-

composition [99],[100] has been applied to turbulent flows in cavities [101], [102]. In structural mechanics, Mottershead and his co-workers [103] have pioneered the use of strain decomposition using Chebyshev and Zernike polynomials [104] to decompose strain fields by treating them as images, which elegantly avoids the difficulty of data existing in arrays with different grids and orientations. The vector and coefficients resulting from such a decomposition process are often known as a feature vector and shape descriptors; and the use of orthogonal decomposition ensures that feature vectors are unique representations of the original data.

One of the areas where feature extraction techniques have been widely employed is structural health monitoring (SHM). Worden et al. [105] extracted frequency response characteristics from the time-series data of a gearbox to identify the onset of a tooth fault. To do so they combined the extracted features with the Mahalanobis-squared distance to determine a threshold that would allow them to identify the onset of damage, assuming normality in the experimental data. Other feature extraction techniques commonly employed in SHM applications include parametric time series methods such as AR, ARMA or ARX models and non-parametric methods through the extraction of statistics either from the raw time-series data or through transformed data such as the fast Fourier transform, wavelet transform or PCA [106],[107], [108]. In addition to the aforementioned techniques, it is not uncommon for researchers to extract multiple response characteristics from the time or time-transformed domains as in [109] to determine the features that maximize the information content that can be subsequently used as damage diagnostics.

Feature-based model validation

After successfully assembling the feature vectors that represent the measured and predicted fields, the question that arises is how these vectors can be used in model validation. For the case of strain fields, Sebastian et al. [7] established a simple decision rule, based on the decomposition of images, using Chebyshev polynomials, for the acceptance or rejection of a model. This is reflected in equation (2.1) where the outcome of the comparison of the feature vectors, i.e. to reject or accept a model as being valid, is based on the uncertainty in the

measurements. Lampeas et al. [9] adopted the method proposed by Sebastian et al.[7], using Zernike polynomials to decompose displacement and strain fields of a beam in three-point bending and proposed the use of a concordance correlation coefficient to calculate the agreement between predictions and measurements.

Dvurecenska et al. [11] developed a metric, based on the relative error between the predicted and measured feature vectors, using Chebyshev or Zernike polynomials. The outcome is a probabilistic statement describing the percentage similarity of the datasets given the measurement uncertainty. The latter was assumed to be spatially constant. The authors suggest that the output of the metric is the probability that the model is representative of reality for a specified intended use, while its calculation is accomplished in three steps: i)initially a relative error e_k is calculated between every measured (S_{M_k}) and predicted (S_{P_k}) shape descriptor which is then normalised by the maximum measured shape descriptor (figure 2.8 middle). This is reflected in equation (2.13).

$$e_k = \left| \frac{S_{P_k} - S_{M_k}}{max_{m \in S_M} |S_{M_m}|} \right| \tag{2.13}$$

Afterwards, each relative error (e_k) is weighted by the sum of the relative errors thus transforming it into percentage (w_k).

$$w_k = \frac{e_k}{\sum_{k=1}^{n} e_k} * 100 \tag{2.14}$$

Finally, the weights w_k of the relative errors whose value is smaller than the error threshold ($e_k < e_{th}$) are summed and produce the outcome of the metric (bottom of figure 2.8). The error threshold (e_{th}) corresponds to the measurement uncertainty in this normalised space and is calculated as:

$$e_{th} = \frac{2u_{exp}}{max_{m \in S_M} |S_{M_m}|} * 100 \tag{2.15}$$

where $2u_{exp}$ is the magnitude of the expanded measurement uncertainty. The output of the validation metric can be mathematically described by equation

(2.16) where || is the indicator function.

$$VM = \sum_i w_i ||_{e_k < e_{th}} \tag{2.16}$$

For the measured dataset of figure 2.2 and the simulated one of figure 2.4 the procedure followed for the calculation of the validation metric is demonstrated in figure 2.8. On the top of the figure the shape descriptors corresponding to the experiment and the simulation are shown. The final validation statement suggested by the authors in the paper should take the following form: 'there is a 100% probability that the model is representative of reality, when simulating y-direction displacements induced by a three-point bending loading of 9.8kN, based on experimental data with 19.2% relative uncertainty'. This outcome is reflected both at the middle and at the bottom of figure 2.8 where the magnitude of all of the normalised errors is less than that of e_{th}, which represents the magnitude of the uncertainty in the feature vector domain. Given the magnitude of the differences between the two fields shown in figure 2.4, it seems that this outcome overestimates the effect of measurement uncertainty in the comparison.

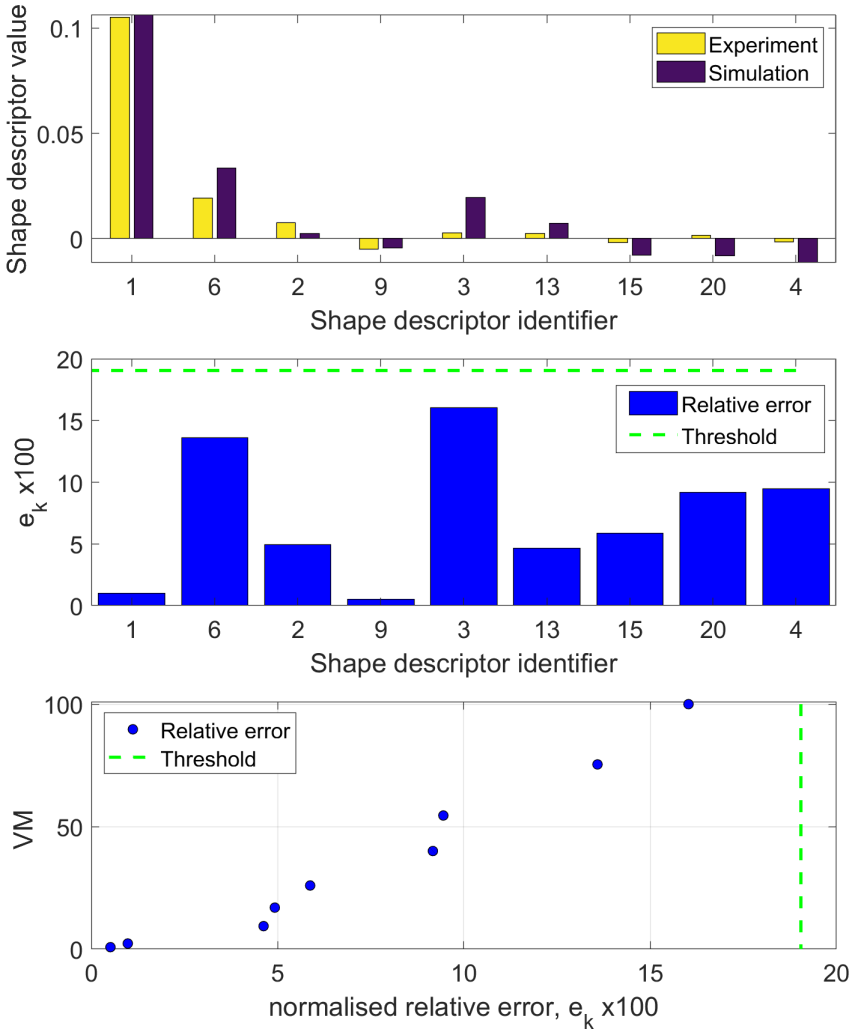

Figure 2.8: The procedure suggested by Dvurecenska et al. [11] for the validation of models via their feature vector form is shown. On top, the shape descriptors corresponding to the measured and predicted datasets corresponding to figures 2.2 and 2.4 respectively are shown. In the middle, the normalised relative error, multiplied by 100, for each shape descriptor is depicted. Finally, in the bottom, the cumulative sum of the weights corresponding to the relative errors whose magnitude is lower than e_{th} is demonstrated. In this scenario the output is 100% for an error threshold of 19.2%

Allemang et al. [10] extracted frequency characteristics across a spatial domain from experimental structural dynamics tests which then assembled into two matrices. Due to lack of simulation outputs they considered one dataset to represent the simulation outcome and the other the measurement. Afterwards they used principal component analysis (PCA) to decompose the two matrices into a number of eigenvalues and eigenvectors. To establish that the first eigenvectors stemming from the different datasets were similar, they calculated the correlation coefficient between the two, and concluded that they were (similar), as its value was on average 0.95. After that, they plotted the eigenvalue progression for the

first principal component for each dataset against each other in a manner similar to that of figure 2.2. The outcome of their metric is the slope corresponding to a fitted line between the two. They also report the associated 'uncertainty' which is characterised by the scatter of points around the fitted line.

Even though this comparison is quantitative, its result is quite baffling. The outcome of validation is the magnitude of the slope, which is difficult to interpret for decision making. Moreover, it is unclear what the scatter about this fitted line represents or how it can be used to inform further actions, given the fact that no precise validation requirements were established. If their suggestions were to be repeated for the data shown in figure 2.4, the resulting slope value of 1 corresponding to the fitted line would suggest that the prediction is a perfect representation of reality, only to be further reinforced by the correlation coefficient between the two which is $\rho = 0.965$. However, the output from the comparison of the datasets depicted in the left side of this figure suggests otherwise. Special care must thus be given when decomposition techniques are used in model validation. Plotting values of coefficients that correspond to components, whose contribution to the underlying dataset is not equal, should be performed carefully.

Li and Lu [110] used PCA to decompose synthetically generated measurements and predictions with multiple responses into a set of coefficients, each corresponding to a principal component and then used the area metric [52] to quantify the proximity of the model to the prediction at a single validation site for each principal component. Then, they extended the notion of u-pooling to assess the global accuracy of the model at multiple validation sites by calculating a weighted sum of the previously acquired values, where each weight is equal to the percentage of variance explained by the respective principal component. Subsequently, Liu et al. [111] applied this technique using measurements and predictions in multiphase flows. Finally, Wang et al. [112] used the Karhunen–Loève expansion to extract significant features from dynamic (time series) measurements and predictions along with the area metric and u-pooling.

2.3 Summary of the review

Following this review, it becomes apparent that model validation has a wide philosophical background, much of which is reflected in probability and statistics. From hypothesis testing, where a model can only be falsified, to Bayesianism where the belief about a model's validity is increased iteratively through empirical successes, model validation is a cross-disciplinary field, concretely associated with every discipline where predictions are used to drive decisions and whose importance is signified by the outcome of those decisions.

In engineering a central question is: 'how good is the model?' A response to that question is provided by a validation metric, capable of comparing the predicted to the measured quantities. Given that nothing is deterministic in the real world and that uncertainties are present during decision making, validation metrics that can account for those uncertainties, in an objective manner, while producing an outcome which is understandable by managers and technical experts alike, should be at the epicentre of research. Depending on the number of response characteristics considered during validation, metrics ranging from univariate to multivariate have been developed across different disciplines attempting to fill the gap in emerging validation practices.

Even though many of these have already been used in applications across disciplines, an enduring problem that practitioners are faced with, is how to select the most appropriate one for the problem at hand and how can its output be evaluated for better decision making. Compared to various metrics such as the Euclidean or Manhattan distance which are simple to understand and have experienced a wide range of applications, the literature on some of the more recent, probabilistic metrics used to assess dissimilarities between distributions is limited. An in-depth explanation of their output, along with a demonstration of their capacity (or lack of) to accurately identify discrepancies between probabilistic predictions and measurements, both in the univariate and the multivariate space will be incorporated in the next chapter. Filling this gap would allow validation practitioners identify the best metric/technique for the problem at hand while understanding its capabilities and limitations.

When it comes to field measurements, which are commonly characterised by

high levels of spatial autocorrelation, and the similarity between fields that are in grids, with different discretization levels must be assessed, feature extraction techniques that reduce the dimensionality of the problem while retaining the significant characteristics of the response provide a viable solution. This is an efficient alternative to geostatistical techniques where limitations arising from various assumptions that must hold true to accurately analyse their similarity hinders their use. One of the knowledge gaps, in validation using feature extraction techniques, is the lack of an accurate representation of the measurement uncertainty in the reduced dimensionality, feature vector space. As demonstrated in figures 2.4 and 2.8, the existing methods used to represent that uncertainty, tend to overestimate its magnitude in that space, potentially deeming bad models valid, whereas simply comparing the two datasets in the initial, spatial domain could lead to a different conclusion. Moreover, the existing techniques assume that the field of uncertainties is spatially constant, an assumption which may not always hold true. This knowledge gap will be resolved in chapter 4, with the aid of approximate Bayesian computation.

Another drawback for most of the existing metrics is the difficulty of communicating their outcome to non-experts, thus transforming them into model selection tools when alternative models are being considered. Even though some of them retain the units of the response quantity, thus making them attractive for engineering applications, their outcome can be still puzzling when the response quantities are transformed into a different space using some decomposition technique. An improvement to this issue is proposed in chapter 5 where a metric based on the pixel-wise differences between two datasets, similar to a reliability metric, can be used to communicate the percentage of differences that can be attributed to the measurement uncertainty. This approach will provide a comprehensive solution to communicating the results of a model validation activity to non-experts.

Chapter 3

Probabilistic model validation

3.1 Introduction

As noted in the introduction and in various verification and validation guidelines, one of the components in establishing credibility in the results of a model lies in the accurate characterisation of the various uncertainties present during a validation process. These uncertainties can be associated with a parameter or constant (e.g. mechanical properties and geometric tolerances), with the quality of measurements (e.g. measurement error) or the limitations of a theory (e.g. bounds on extrapolation).

A wide range of methods capable of characterising the various forms of uncertainty have been developed. These include probability theory [41], fuzzy set theory [113], interval analysis [114], p-boxes [115] and Dempster-Shafer theory [116] among others. In the case of probability theory these uncertainties are characterised using probability distributions and may represent parameter variability or epistemic uncertainty (e.g. measurement error). Following their characterisation, the propagation of these uncertainties through a model can be easily performed using a technique known as a Monte Carlo simulation [117].

During a Monte Carlo simulation, samples are drawn from the distributions characterising the various uncertainties which are then used as inputs in the model. The result after running the model across the different parameter combinations is a distribution of response outputs that reflect the uncertainty in their

inputs. On the experimental side, multiple experiments are similarly executed in order to acquire an accurate representation of the associated uncertainties.

Having acquired the data from the simulation outputs and the measurements, the objective is to accurately characterise the capability of a model to represent the real world for the given uncertainties. In this chapter, the various metrics capable of assessing probabilistic model outputs will be reviewed. The review will start from two extensively cited univariate metrics which are the area metric and u-pooling and will then extend to multivariate ones such as the Mahalanobis distance area metric and the probability integral transform area metric. The associated theory will be described and a series of numerical and engineering examples will be employed to demonstrate their capabilities and limitations. The results of these examples can be used as a guide for probabilistic model validation and will set the basis of the research described in subsequent chapters.

3.2 Background theory and metrics

3.2.1 Empirical distribution functions

It is normal practice for experimentalists to obtain multiple measurements (when possible) that allow them to draw inferences about a quantity of interest. Then, using some descriptive statistic or confidence intervals, the population parameters from which these values are assumed to arise from can be estimated. One way to present multiple observations non-parametrically is to use empirical distribution functions (EDF) [41]. They represent the probability of measurements (or simulations) being equal or smaller than a given value t; they monotonically increase and range between [0,1] (the probability space). Traditionally, empirical distribution functions are used to identify whether the empirical data are adequately represented by the selected family of probability distributions (e.g. normal, uniform, etc.) or to quantify the difference between various datasets (e.g. model predictions vs experimental measurements). Their mathematical formula is given below:

$$\widehat{F_n}(t) = \frac{\text{number of measurements} \leq \text{threshold}}{\text{total number of measurements}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{x_i \leq t} \tag{3.1}$$

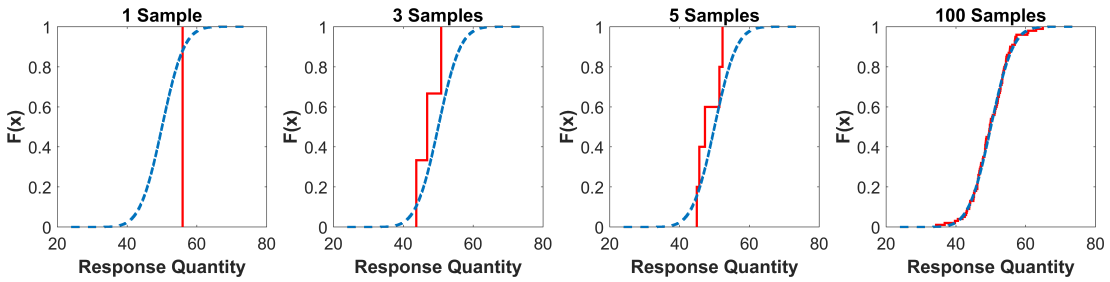

Figure 3.1: Empirical distribution function (step curve) against the cumulative distribution function (continuous curve) of a normally distributed variable.

The $\widehat{F_n}$ represents the fact that the empirical distribution function is an estimator of the cumulative distribution function; n is the number of measurements that the sample comprises of and $\mathbf{1}$ corresponds to the indicator function. The indicator function is a function whose value is unity within the domain of the function (in this case for $x_i \leq t$) and zero elsewhere.

Figure 3.1 demonstrates examples of the shape of empirical distribution functions of a normally distributed variable (or response quantity) as the number of measurements increases. It is obvious with a greater number of measurements the empirical distribution function becomes more smooth and in the limit of $n \to \infty$ the empirical distribution function completely overlaps the cumulative distribution function (CDF). The latter is the continuous analogue of the empirical distribution function when the probability distribution of the random variable is known. It should be stated that the terms empirical distribution function and distribution function (DF) are interchangeably used within the text.

3.2.2 Area metric

Having explained how distribution functions can be used to summarize model predictions or experimental measurements, the next step is to determine a method to quantify their differences. One of the metrics available to achieve that is the area metric. The area metric is a measure of statistical mismatch between two empirical distributions which in this case consist of the experimental against the predicted. This is done via the calculation of the integral of the absolute difference between the two and can be mathematically described as:

$$A = \int_{-\infty}^{+\infty} |F(x) - G(x)| dx \tag{3.2}$$

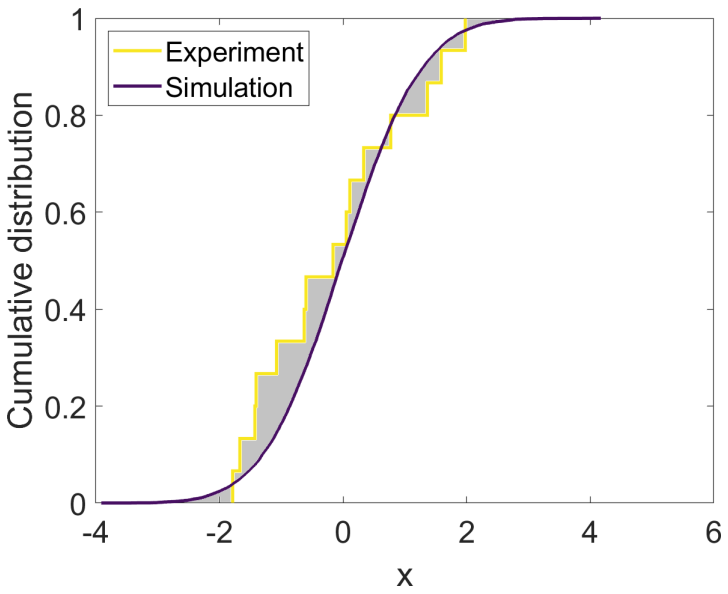

Figure 3.2: Empirical distribution functions corresponding to the measurements and the predictions. The predictions are depicted by the continuous purple line, while the measurements by the yellow stepped line. The grey area between them constitutes the output of the area metric.

where: $F(x)$ is the measurements' distribution function and $G(x)$ is the model predictions' distribution function. A demonstration of the area metric is given in figure 3.2 where the outcome of the comparison between the two distribution functions is described by the grey area in-between. Moreover, 95% confidence intervals will be calculated for the area metric where appropriate. The calculation of these intervals is based on bootstrap samples [118] where a total of 10000 bootstrap samples were taken for each case.

3.2.3 U-pooling

Even though the area metric allows empirical distributions to be compared in a single validation site, it does not answer the question of how can one quantify the validity of the model across multiple sites. A validation site can be considered as a point in the input space where the validation experiment is executed. The validation site consists of parameters such as loading level, boundary conditions or temperature which are known as control parameters [3] and uniquely characterise it. It would be desirable if model-measurement comparisons across validation sites could be collected in a simple and efficient manner; for example to jointly assess the capability of a model to predict linear and non-linear responses after certain

loading levels. An answer for the case of probabilistic modelling emerged by Ferson et al. [52] in the form of u-pooling.

The steps needed to implement u-pooling will be described and are shown in figure 3.3. Initially each of the p experimental measurements $\{y_p^e\}$ are summarized using the empirical distribution function described earlier. Then the same process is repeated for the q model outcomes $\{y_q^m\}$. For most of the cases $p \ll q$ due to the high costs involved in experimental testing. Afterwards, making the assumption that the experimental measurements are following the same distribution as the computational prediction, the experimental measurements are transformed into u-values using the model's DF. For each of the of the p measurements the transformation is portrayed by the equation $u_i = F^m(y_i^e)$ (i is the identifier for each separate measurement while F^m represents the model's DF). This process is repeated until all of the experimental measurements have been transformed in the respective u-values (universal probabilistic scale). Afterwards, the distribution function of the u-values is calculated and plotted against the CDF of a uniformly distributed variable $U(0, 1)$. The reason for this lies in the use of the probability integral transform (PIT) [119] according to which if random variable Y is defined through the CDF $F(X)$ of a continuous random variable X, then Y is uniformly distributed in the domain $[0, 1]$ and its CDF takes the form of a 45° line in the same domain. Any discrepancy between the two distribution functions would be portrayed in the deviation from the 45° line. Possible deviations can be then quantified using the area metric between the two newly formed DFs. A flowchart of the procedure is portrayed in figure 3.3. It should be stated that even though the predictions' distribution function acts as the reference in this case and later, this constitutes a numerical simplification. It is usually easier to obtain multiple numerical simulations rather than perform multiple experimental tests, allowing a continuous distribution to be determined against which the measured one is assessed.

An example is given to visually describe the process. Three measurements (strain) are pooled together. The three vertically arranged graphs of figure 3.4 depict the first steps of the procedure where each of the three measurements is transformed via the model's DF into the respective u-value. Then the empirical

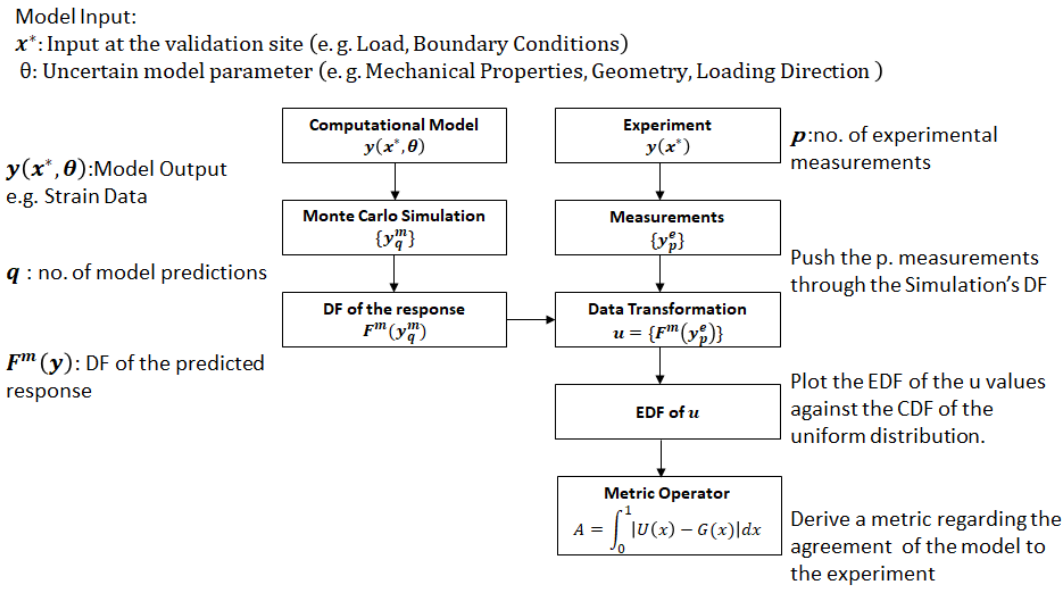

Figure 3.3: U-pooling flowchart.

distribution of those u-values is calculated (middle, $G(x)$). The CDF ($U(x)$) of a uniformly distributed variable $U(0, 1)$ is overlaid against the $G(x)$ in the right graph. Finally, the grey area representing the difference between the two curves is calculated using the area metric. It should be noted that in this example the model's DF is the same across the three measurements, implying that the measurements were taken at a single validation site. It becomes apparent then that u-pooling can be used to assess the deviations between two DFs at a single site.

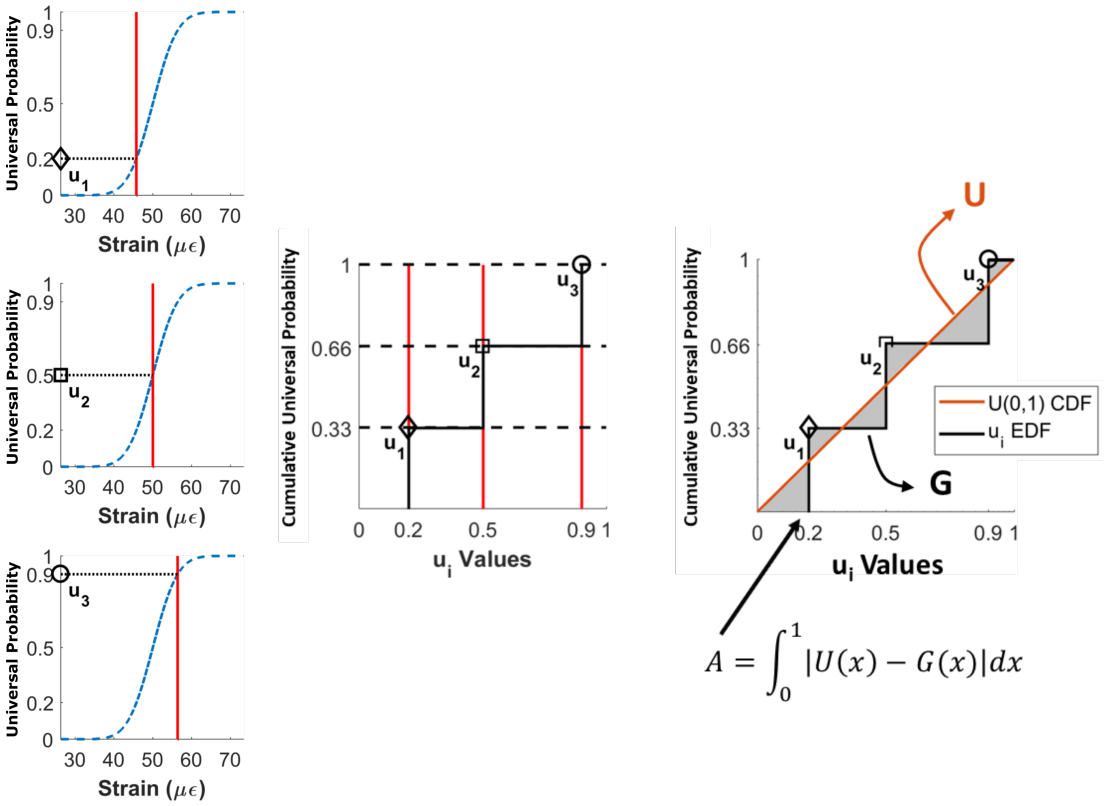

Figure 3.4: U-pooling procedure schematic. Each measurement (shown in red vertical lines) is initially transformed to its corresponding u-value at the point of intersection with the predicted distribution function (continuous blue curve). Afterwards, the empirical distribution function of the u-transformed measurements is constructed (middle) and the area metric between this distribution function and the cumulative distribution function of a uniform distribution is calculated (right). Any deviations between the model predictions and the measurements is reflected in the area metric value

3.2.4 Probability integral transform area metric

The need to assess multivariate probabilistic simulations led to the development of the probability integral transform (PIT) area metric by Li et al. [60] which can be considered an extension of u-pooling into the multivariate space for a single validation site. Analytically, there may be cases where certain requirements entail the need for the validation procedure to be performed using multiple response quantities. These may consist of displacements and deformations at certain locations or depending on the problem may expand to temperature, electric current or acceleration, thus resulting in a multivariate feature vector. The implementation of u-pooling to higher dimensions, where each dimension corresponds to an element of that vector, known as the PIT area metric is based on the use of the

multivariate probability integral transform. The steps needed to implement this multivariate metric are described below and are accompanied by the flowchart of figure 3.5.

- Step 1: The computational model is executed at a specific validation site x^* for a series of uncertain model parameters θ. At the same time, measurements are taken from the experiment. The only difference to u-pooling is that multiple response quantities are simultaneously recorded.

- Step 2: The results of the previous step constitute the simulation outputs: $y^m_{i,k}$ (m corresponding to model) and the measurements: $y^e_{i,j}$ (e corresponding to experiment) . The index i corresponds to the i_{th} component of the feature (representing displacement, acceleration or some other response quantity), while the indexes k and j correspond to the k_{th} and j_{th} simulation and measurement respectively. The dimensionality of the problem is d, while the number of simulation outputs and measurements is q and p respectively.

- Step 3: The multivariate empirical distribution function F^m is built from the simulation outputs. The value of F^m for each of the simulation outputs is designated as Y_k. The measurements are transformed into v-values using F^m. This converts them into probabilities in a similar manner as u-pooling, the only difference being that the model's distribution function is now multivariate.

- Step 4: The construction of the empirical distribution K^m using the Y_k values from the previous step takes place. The newly built distribution function is based on the multivariate probability integral transform which means that instead of a 45 ° line identified as U in figure 3.4 now the resulting curve is distinctive as it reflects certain characteristics of the underlying multivariate distribution function such as correlation. At the same time the corresponding distribution function S^e from the v-transformed measurements is constructed and plotted against the K^m

- Step 5: The quantitative comparison between the two distribution functions takes place. This is done with the aid of the area metric. If the two are

perfectly aligned then this value will be zero; otherwise a positive value of the area metric will be indicative of deviations.

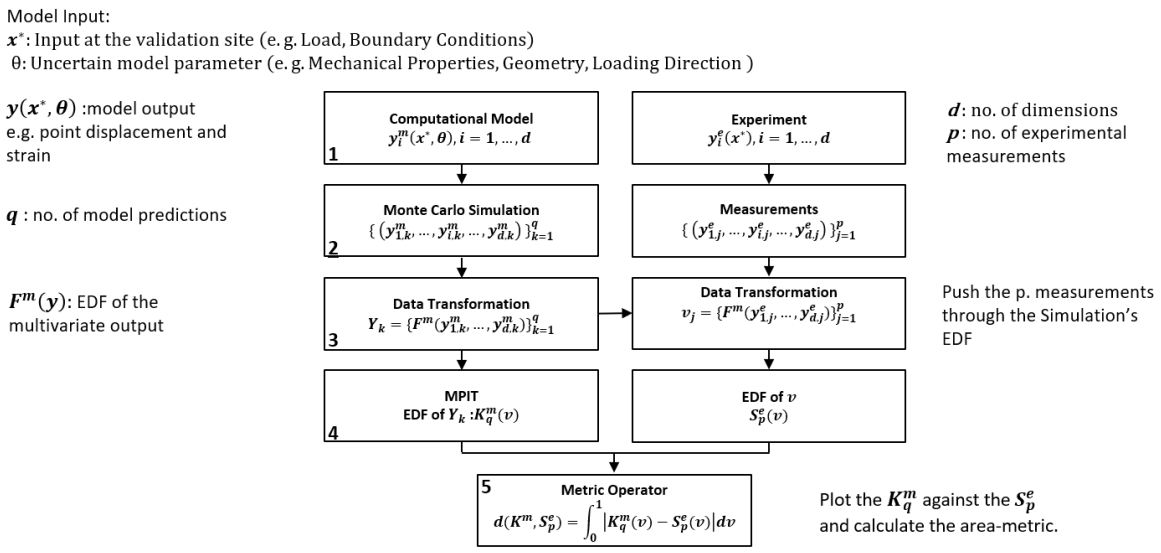

Figure 3.5: PIT area metric flowchart

A visual explanation of the process is given in figure 3.6. The data consist of displacement and strain measurements and predictions resulting from a numerical analysis of a simply supported beam. Details of the datasets, both of which were numerically generated, will be given later. Having thus acquired a total of 100 predictions and 10 measurements the results are plotted on the top left. Afterwards, using the model outputs the two-dimensional distribution function is built as shown on the top right. The vertical dashed lines represent the measurements in this space whose cross-sections with F^m constitute of their v-values. Subsequently the model predictions are transformed into the K^m distribution function as shown in the 90° anti-clockwise rotated figure in purple. At the same time the measurements are transformed in a similar manner to u-pooling into the S^e distribution function. Finally, the two distributions are plotted against each other and the area metric between the two is calculated. The range of v values in the x-axis of the same figure reveals that both measurements and predictions have covered less than the 60th percentile of the prediction's distribution function. In order to achieve a wider range of v-values a larger number of simulations is needed. This phenomenon is only aggravated as the dimensionality of

the problem increases and more simulations and measurements are required to adequately cover the whole space.

Generally, as the dimension of the input space increases, so does the need for acquiring (exponentially) more samples to draw valid inferences, an issue that can prove to be very costly when either a large number of simulations or experiments is needed to draw inferences of a certain precision. Trying to tackle this issue, known as the 'curse of dimensionality' [120], numerous developments have emerged: techniques such as Latin Hypercube sampling [121] stratify the input space achieving more precise estimates compared to others (e.g. random sampling) for the same number of samples. Markov chain Monte Carlo techniques [46] allow samples to be drawn from high-dimensional distributions and have set the basis for many developments in the field of computational statistics. Both techniques will be used in this and subsequent chapters to enable high-dimensional comparisons between measurements and predictions.

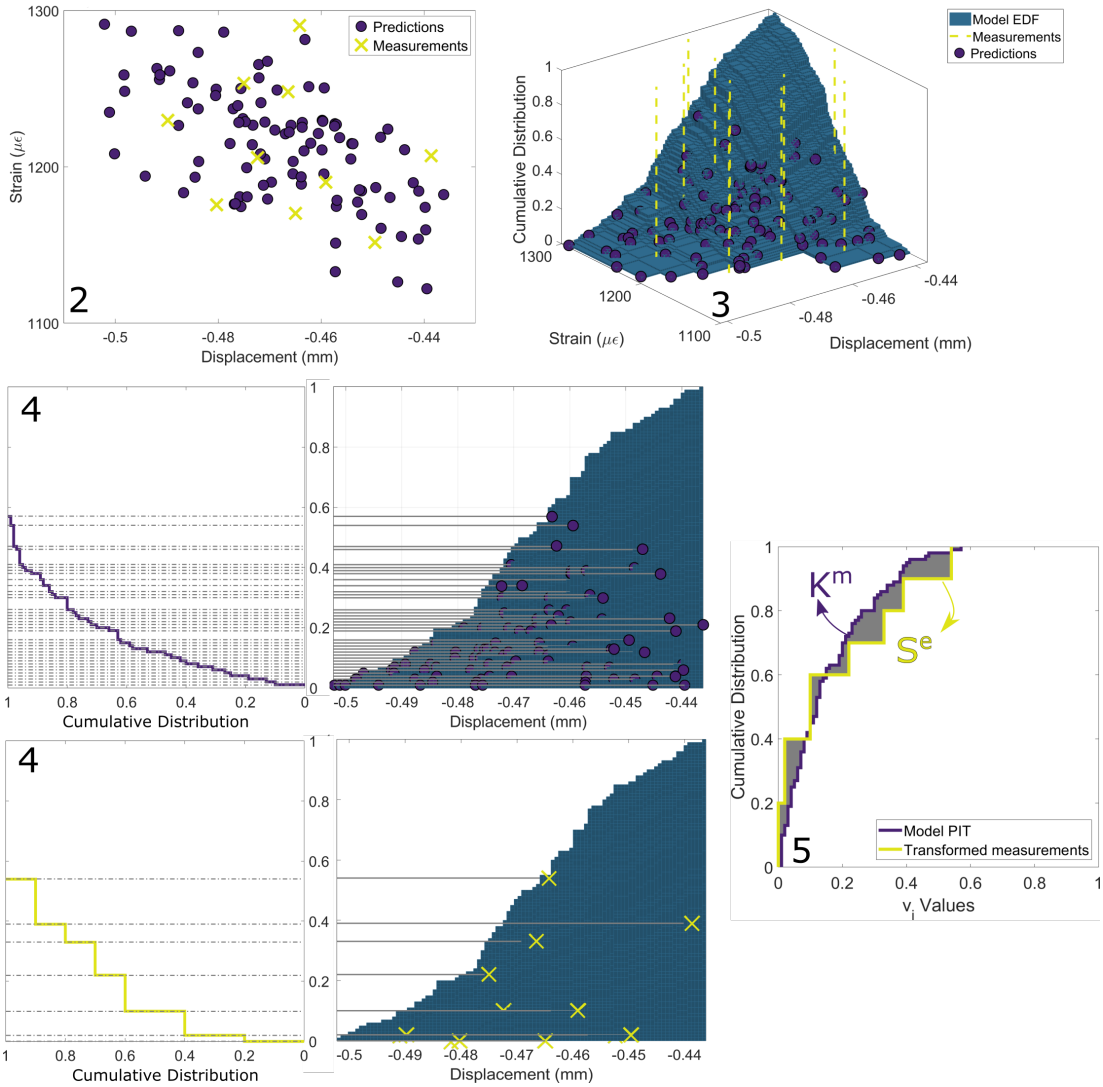

Figure 3.6: PIT area metric procedure schematic. On the top left the measurements are shown against the predictions (2). In a process similar to u-pooling the experimental measurements (yellow dashed lines) are transformed into v-values using the model's distribution function (4). The latter is constructed by the model's outputs (purple circles) (3). Compared to the 1-D case where the model's probability integral transform results into a continuous 45° curve, in higher dimensions it follows a distinctive pattern (middle). Finally, the area metric between the two newly-formed curves is calculated (shown by the grey area)(5).

3.2.5 Mahalanobis distance area metric

The Mahalanobis distance (MD) area metric [66] is the last method that will be used to assess the accuracy of probabilistic model predictions. As described in the literature review the Mahalanobis distance can be considered an extension of the Euclidean distance to account for any prevalent uncertainty via its normalisation by the covariance matrix. Analytically, the Mahalanobis distance

can be characterised as a normalised measure of distance of a point, in this case corresponding to a measurement, from the mean of a distribution, corresponding to the probabilistic output of a model.

This means that the Mahalanobis distance quantifies the distance between a point (single multivariate measurement) and a distribution (collection of simulation outputs) and not between two distributions which is desired. To bypass this issue, the technique developed by Zhao et al. [66] will be used. Their technique can be described in three steps; in the first step the multivariate measurements and predictions are assembled for a specific validation location as in the PIT area metric. Afterwards, the Mahalanobis distances of each measured quantity are calculated against the prediction outcomes. This results in a distribution of 'experimental' Mahalanobis distances. At the same time, the MDs of each of the prediction outcomes is calculated with respect to the multitude of the prediction outcomes. This results in a distribution of 'simulated' MDs which is then compared to the experimental one using the area metric. These steps are depicted in the flowchart of figure 3.7.

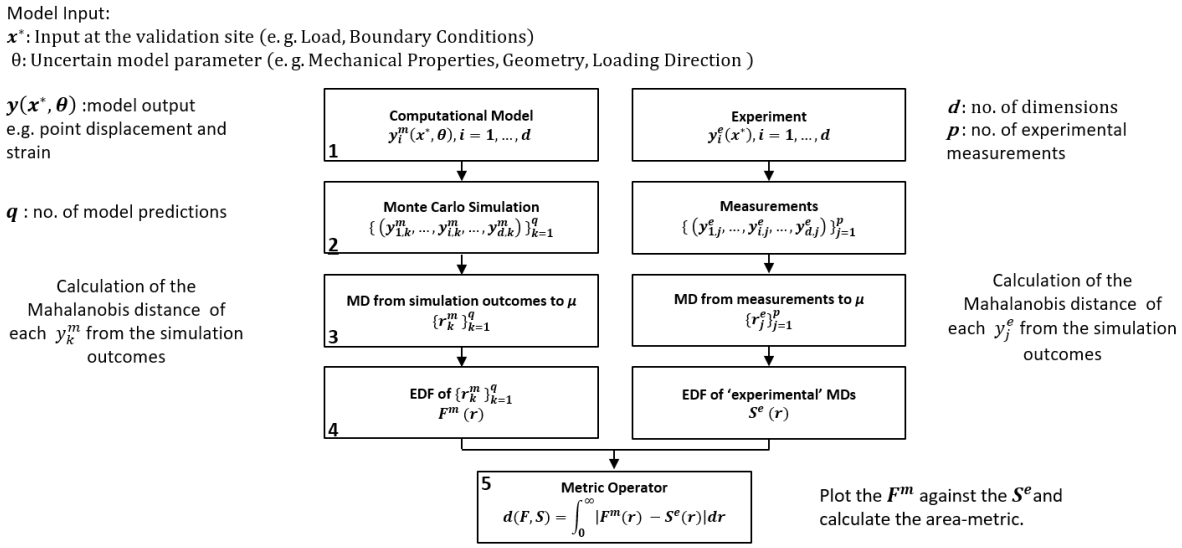

Figure 3.7: Mahalanobis distance area metric flowchart.

To assist the explanation of the Mahalanobis-based area metric, figure 3.8 will be used. On the left side of the figure one can see the predictions and measurements as in figure 3.6 with the addition of contours corresponding to equidistant

loci from the mean of the predictions. It is obvious that these loci form ellipses whose orientation is aligned with the orientation of the simulation outcomes. On the right side, the empirical distribution of the simulations' Mahalanobis distances is shown against the respective one from the measurements. The magnitude of the gray area between the two is calculated using the area metric. Compared to previous figures where the x-axis values corresponded to probabilities and thus ranged between 0 and 1 , now the values on the x-axis depend on the Mahalanobis distance of the points (predictions or measurements) from the mean of the predicted outcomes thus having no specific upper bound.

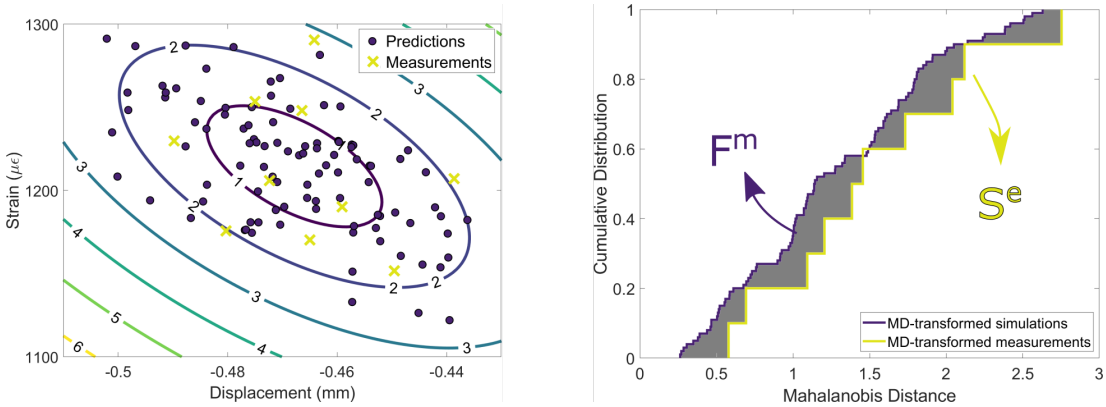

Figure 3.8: Mahalanobis distance area metric procedure schematic. On the left graph isocurves corresponding to loci of equal Mahalanobis distance are shown. Each prediction is transformed based on its Mahalanobis distance from the distribution consisting of the various simulation outputs. The empirical distribution corresponding to the Mahalanobis distance-transformed predictions is shown in purple in the right graph. The empirical distribution of Mahalanobis distance-transformed measurements is shown in yellow.

3.2.6 Full-field data decomposition using modified Chebyshev polynomials

In this section the basic mathematical background regarding the decomposition of full-field data using a predefined number of modified Chebyshev basis functions will be demonstrated. One of the methods used to extract features from data (and thus represent them in a lower-dimensionality space) is through their moments. The general two-dimensional moment definition using a weighting function ψ_{pq}

and an image intensity function $f(x, y)$ is given by equation 3.3 [122].

$$\Psi_{pq} = \int_{-1}^{1}\int_{-1}^{1} \psi_{pq}(x, y) f(x, y) dxdy \qquad p, q = 0, 1, 2, .. \tag{3.3}$$

For example, the weighting function used to extract geometric moments of order $(p + q)$ on an image of size $N \times N$ pixels is defined as

$$\psi_{pq}(x, y) = x^p y^q, \quad 0 \leq x, \quad y \leq N - 1 \tag{3.4}$$

However, the issue with this type of moment generation is that they suffer from large variation in the dynamic range of values for different orders of the polynomials and can thus cause numerical instabilities. To avoid such issues, orthogonal polynomials are selected. Analytically, consider a system $\{f_n(i)\}$ where $a \leq i \leq b$. The orthogonality property is then written as

$$\sum_{i=a}^{i=b} w(i) f_m(i) f_n(i) = \rho(n, a, b)\delta_{mn} \tag{3.5}$$

where $w(i)$ is the weighting function, $\rho(n, a, b)$ is the squared norm, δ_{mn} is Kronecker's delta and m, n correspond to the order of the polynomial. For the case of Chebyshev polynomials the weighting function is the unity. The reason behind the selection of Chebyshev polynomials to represent the data is their faster convergence rate to the initial values compared to Legendre polynomials for the same order and the fact that they can be used in a rectangular grid without having to transform them in a new coordinate system (e.g. unit circle for the case of Zernike polynomials). The respective discrete Chebyshev moments are then given by equation 3.6.

$$T_{pq} = \frac{1}{\rho(p, N)\rho(q, N)} \sum_{x=0}^{N-1}\sum_{y=0}^{N-1} t_p(x) t_q(y) f(x, y) \qquad p, q = 0, 1, 2, ...N - 1 \tag{3.6}$$

where p, q are the orders of the polynomials. The original Chebyshev polynomials are defined as:

$$t_n(x) = n! \sum_{k=0}^{n} (-1)^{n-k} \binom{N-1-k}{n-k} \binom{n+k}{n} \binom{x}{k} \tag{3.7}$$

and

$$\rho(p,N) = (2n)!\binom{N+n}{2n+1}, \qquad n = 0,1,....,N-1 \tag{3.8}$$

However, the original Chebyshev polynomials can lead to numerical instabilities as the number of data increases. For this case Mukundan et al. [122] developed the scaled polynomials as:

$$\tilde{t}_n(x) = \frac{t_n(x)}{\beta(n,N)} \tag{3.9}$$

while the updated moments are given by the equation:

$$T_{pq} = \frac{1}{\tilde{\rho}(p,N)\tilde{\rho}(q,N)} \sum_{x=0}^{N-1}\sum_{y=0}^{N-1} \tilde{t}_p(x)\tilde{t}_q(y)f(x,y) \qquad p,q = 0,1,2,...N-1 \tag{3.10}$$

where:

$$\tilde{\rho}(n,N) = \frac{\rho(n,N)}{\beta(n,N)^2} \tag{3.11}$$

and the inverse transformation is:

$$f(x,y) = \sum_{m=0}^{N-1}\sum_{n=0}^{N-1} T_{mn}\tilde{t}_m(x)\tilde{t}_n(y) \qquad x,y = 0,1,...,N-1 \tag{3.12}$$

$$\beta(n,N) = N^n \tag{3.13}$$

Then the scaled Chebyshev polynomials are given by

$$\tilde{t}_n(x) = \frac{(2n-1)\tilde{t}_1(x)\tilde{t}_{n-1}(x) - (n-1)(1-\frac{(n-1)^2}{N^2})\tilde{t}_{n-2}(x)}{n} \qquad n = 2,3,...N-1 \tag{3.14}$$

and

$$\tilde{\rho}(n,N) = \frac{N(1-\frac{1}{N^2})(1-\frac{2^2}{N^2})...(1-\frac{n^2}{N^2})}{2n+1} \qquad n = 2,3,...,N-1 \tag{3.15}$$

The first fifteen two-dimensional Chebyshev kernels are depicted in figure 3.9.

As stated in equation 3.15 the maximum number of kernels that can be used to recreate the data field is limited by the number of data points. Of course this sets an upper bound regarding the reconstruction process. However, most of the time only a small amount of kernels is required to accurately reconstruct the

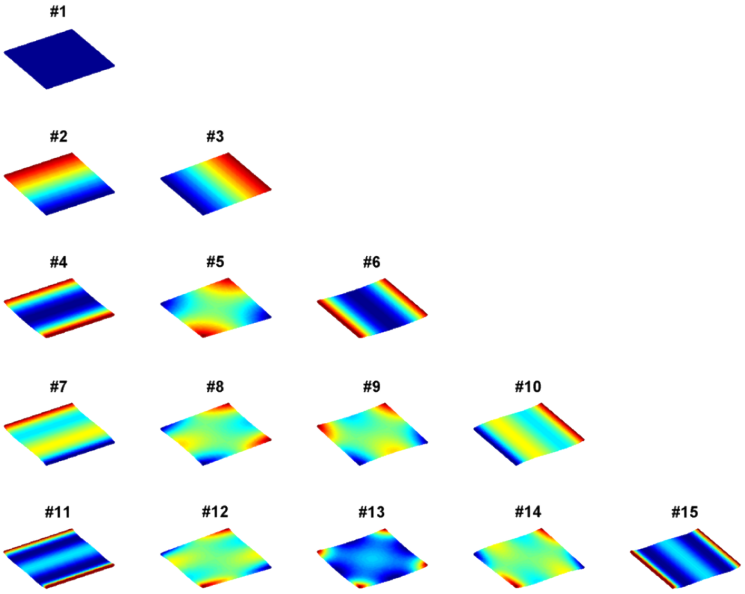

Figure 3.9: First fifteen 2D Chebyshev shape descriptors [123].

initial data (ranging from just a few to the order of tens related to the smoothness of the underlying field) .

3.3 Numerical examples

A series of numerical examples will be presented in this section. These will be employed to demonstrate the effects of various parameters such as differences in the means, standard deviations and sample sizes between the predicted and measured quantities on the described metrics. The section will be divided into two parts, each associated with the dimensionality of the problem. The first will focus on 1-D examples and the use of the area metric and u-pooling, whereas the second will expand on higher dimensionality problems while employing the PIT area metric and the Mahalanobis distance area metric.

3.3.1 1-D examples

14 examples will be employed to demonstrate the effect of varying parameters on the area metric and u-pooling. The parameters that characterise the two distributions, one corresponding to the experimental measurements and the other to the model's predictions, can be seen in table 3.1. The first column represents the

Table 3.1: Numerical examples' (univariate) parameter definition.

No.	$\mu_{exp}(\mu\epsilon)$	$\mu_{sim}(\mu\epsilon)$	$\sigma_{exp}(\mu\epsilon)$	$\sigma_{sim}(\mu\epsilon)$	N_{exp}	N_{sim}	area metric ($\mu\epsilon$)	area metric 95% CI	u-pooling
1	150	150	3	3	1000	1000	0.32	[0.16, 0.58]	0.03
2	150	150	3	3	6	1000	0.97	[1.00, 2.49]	0.08
3	150	150	3	6	1000	1000	2.75	[2.50, 3.02]	0.12
4	150	150	3	6	6	1000	3.43	[3.18, 4.79]	0.15
5	150	150	3	12	1000	1000	7.81	[7.34, 8.29]	0.18
6	150	150	3	12	6	1000	8.54	[8.07, 9.91]	0.20
7	150	150	3	24	1000	1000	17.96	[17.04, 18.89]	0.22
8	150	150	3	24	6	1000	18.80	[18.01, 20.43]	0.23
9	147	150	3	3	1000	1000	3.27	[3.01, 3.53]	0.27
10	147	150	3	3	6	1000	3.25	[2.12, 4.88]	0.30
11	144	150	3	3	1000	1000	6.27	[6.01, 6.54]	0.43
12	144	150	3	3	6	1000	6.16	[4.68, 7.80]	0.45
13	142	150	3	3	1000	1000	8.28	[8.01, 8.54]	0.47
14	142	150	3	3	6	1000	8.03	[6.60, 9.77]	0.49

μ_{exp}	mean value for the experimental dataset
μ_{sim}	mean value for the simulated dataset
σ_{exp}	standard deviation for the experimental dataset
σ_{sim}	standard deviation for the simulated dataset
N_{exp}	number of experimental measurements
N_{sim}	number of simulation outputs
area metric	area metric calculation result
area metric 95% CI	95% confidence interval for the area metric
u-pooling	u-pooling calculation result

identifier that will be used to refer to each numerical experiment; columns 2-5 outline the parameters of the distributions describing the experiment and simulation; columns 6-7 outline the number of samples drawn from each distribution and columns 8-10 comprise of the outcome of the comparisons using the area metric and u-pooling. The data used across the examples was generated by a pseudorandom number generator in MATLAB.

Figures 3.10,3.12,3.13 depict some of the results of table 3.1 to avoid visual clutter. The figures for the rest of the examples are shown in Appendix A. The left graphs represent the probability distributions that the pseudo-measurements and simulations were taken from. The middle graphs depict the distribution functions from the simulation and the experiment, while annotating the area metric from the comparison of the two and finally, the right graphs reflect the outcome of u-pooling from the comparison.

Figure 3.10 demonstrates the effect of sample size on the area metric and u-pooling. Even though both datasets stem from the same distribution (shown on the left), the area metric value is not zero, an issue caused by the discrete

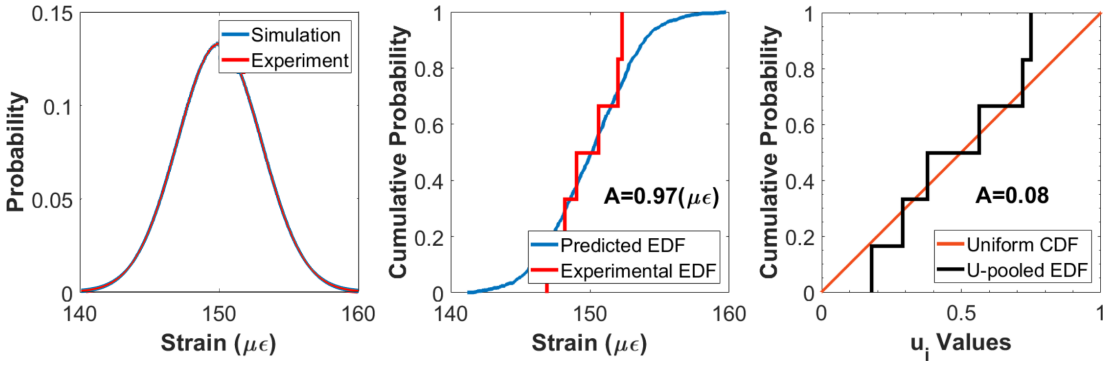

Figure 3.10: Example 2: measurements and predictions follow the same distribution. A nonzero validation outcome, shown across graphs by the A value arises due to the small number of measurements.

nature of their distribution functions. Techniques such as p-boxes [115] and the Dvoretzky-Kiefer-Wolfowitz inequality [124] can be used to determine bounds on EDFs when their sample size is small. However, the use of such techniques is not straightforward when it comes to quantitatively comparing two distributions. In such cases, bootstrapping can be used to determine confidence intervals for the associated metric. The bootstrapped predictions and measurements used to determine the confidence intervals for this example are shown in figure 3.11. This figure reinforces the fact that when a small number of samples is available it is impossible to obtain an area-metric value of 0. However, the resulting confidence intervals could provide decision makers with information regarding the range of outcomes that could be expected. Similar results emerge in example no 1. even though the number of measurements is significantly larger. The next figure (3.12) shows the effect of differences in the means of the distributions. The value of the area metric substantially increases from the previous idealized cases, while the curve of the u-transformed measurements evidently deviates from the ideal 45°line. Given that the range of outcomes in u-pooling is bounded between 0 and 0.5, one may consider this to be a bad model. Finally, figure 3.13 demonstrates the effect of differences between the variances of the experimental and simulated data when their means are equal. In this case the third graph shows that the experimental values lie between the 40th and 60th percentile of the corresponding simulated ones (abscissa).

Even though a series of numerical examples have been employed to demonstrate how the area metric and u-pooling can be used to assess the similarity

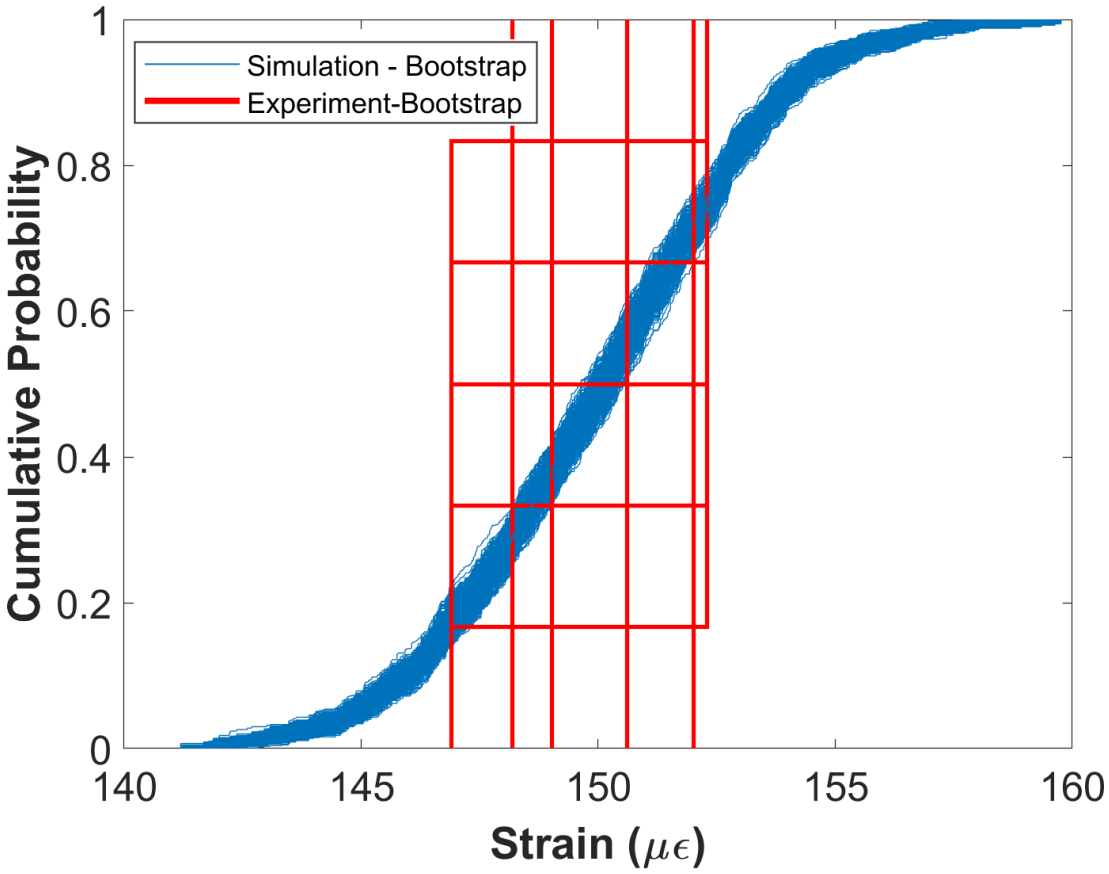

Figure 3.11: Example 2: The bootstrapped EDFs corresponding to the simulations (shown in blue) and the measurements (shown in red) respectively.

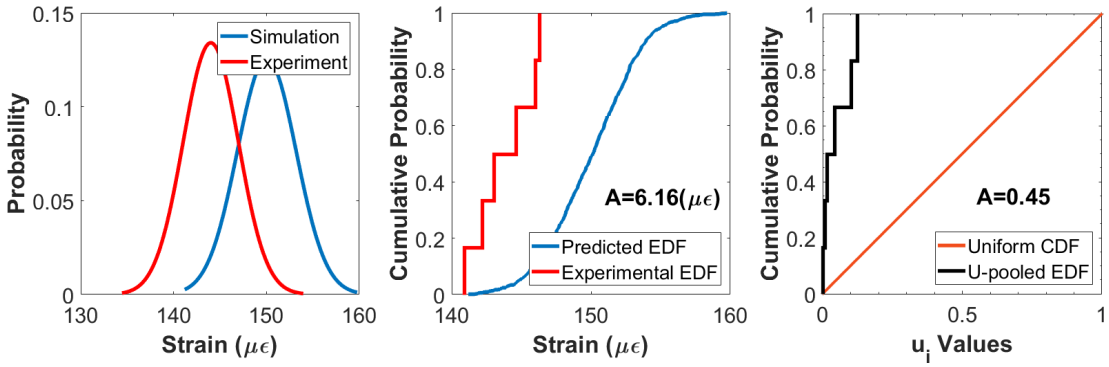

Figure 3.12: Example 12: the mean value of the distributions is different, while the standard deviation is the same.

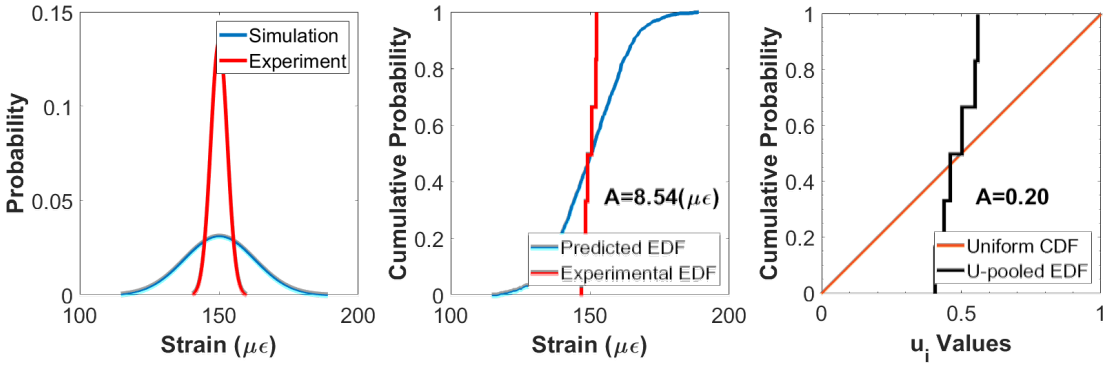

Figure 3.13: Example 6: The mean value of the distributions is the same, while the standard deviation differs.

between two distributions, an extensive analysis was additionally performed to improve the understanding of the effects of the various parameters on the metrics. A total of 100,000 parameter combinations were used during the analysis, each one corresponding to variables such as the means and standard deviations of the measurements and predictions. Each experimental distribution was built using 100 samples while each predicted distribution was built using 1000 samples, both following the normal distribution. The results of the parametric analysis can be seen in figure 3.14.

Visualising 5-dimensional data can be quite bothersome (4 used for the parameters $\mu_{exp}, \mu_{sim}, \sigma_{exp}, \sigma_{sim}$ and 1 for the outcome of the comparison). In order to avoid this problem, the distributions' parameters were transformed so that the x and y axes in the 3-D plots of figure 3.14 correspond to differences in the means and standard deviations respectively. To provide a more detailed explanation of how the two metrics behave, two paths along the 3-D plots have been selected and overlaid in the same plot. The first path shown in purple rhombi corresponds to values where the differences in the means are zero thus allowing for a better understanding of the effect of standard deviations. The second path is depicted using green squares and can be used to assess the effect of means when the difference between the distributions' standard deviations is zero. It can be seen that the effect of the means is slightly larger than that of standard deviations for the case of the area metric as it can be qualitatively assessed by the slopes of the two paths. On the other hand, it seems that the value of u-pooling is more complex as shown in the top right side of the same figure. For the case of $\Delta\sigma_{exp-sim} = 0$

(green squares) it seems that it is influenced by the difference in the means of the two distributions and less from the difference in standard deviations when $\Delta\mu_{exp-sim} = 0$ (purple rhombi).

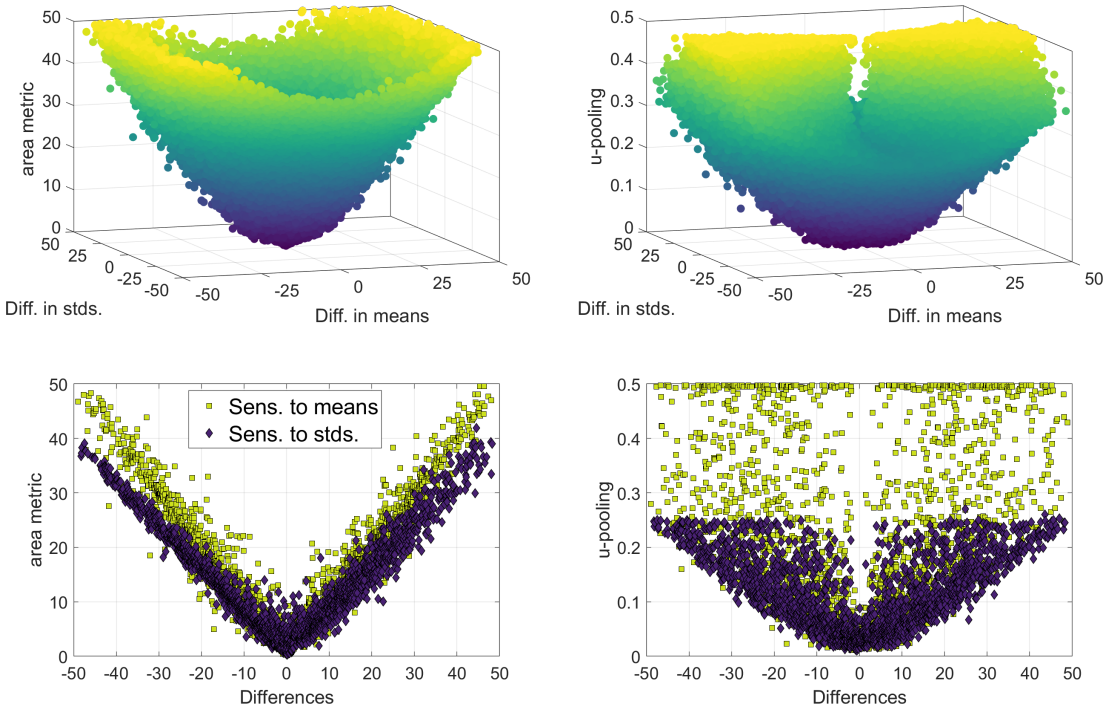

Figure 3.14: The effect of differences in the means and standard deviations of two univariate Gaussian distributions on the area metric and u-pooling respectively is shown on top. At the bottom, the purple rhombi and green squares correspond to points where the differences in the means and standard deviations respectively are zero.

3.3.2 2-D examples

Having demonstrated the use of area metric and u-pooling for the comparison of 1-D distributions, the focus will be shifted to multivariate problems where the accuracy of a model's predictions is assessed jointly on two or more quantities of interest. The objective is to demonstrate how the PIT area metric and the Mahalanobis distance area metric can be used to assess multivariate forecasts, to identify their strengths and limitations and to establish how they can be used for efficient decision making.

In a manner similar to the previous section, a number of numerical examples have been employed. These comprise a series of simulations where parameters including the means, standard deviations and correlation coefficient vary within pre-defined ranges. The results for some of these simulations are shown in figures

3.15,3.16 and 3.17 while the parameter values are outlined in table 3.2. Similar to the 1D numerical examples, the figures corresponding to the rest of the numerical examples are shown in Appendix B. Each figure consists of three graphs; the first depicts the measured and predicted datasets, the second visualizes the calculation of the MD area metric and the third illustrates the calculation of the PIT area metric.

Table 3.2: Parameters and results for the 2-D numerical examples.

	$\mu_{exp_1}(mm)$	$\mu_{exp_2}(\mu\epsilon)$	$\mu_{sim_1}(mm)$	$\mu_{sim_2}(\mu\epsilon)$	$\sigma_{exp_1}(mm)$	$\sigma_{exp_2}(\mu\epsilon)$	$\sigma_{sim_1}(mm)$	$\sigma_{sim_2}(\mu\epsilon)$	ρ_{exp}	ρ_{sim}	MD metric	PIT metric
1	-0.47	1214	-0.47	1214	0.0160	36	0.0160	36	-0.59	-0.59	0.24	0.05
2	-0.46	1244	-0.47	1214	0.0160	36	0.0160	36	-0.59	-0.59	9.31	0.44
3	-0.46	1244	-0.47	1214	0.0320	36	0.0160	36	-0.59	-0.59	9.17	0.30
4	-0.46	1244	-0.47	1214	0.0160	36	0.0160	36	0	-0.59	9.12	0.44
5	-0.48	1184	-0.47	1214	0.0160	36	0.0160	36	-0.59	-0.59	9.50	0.15
6	-0.48	1184	-0.47	1214	0.0160	36	0.0160	36	-0.80	-0.59	9.43	0.15
7	-0.48	914	-0.47	1214	0.0160	36	0.0160	36	-0.59	-0.95	28.40	0.06
8	-0.47	1814	-0.47	1214	0.0160	36	0.0160	36	-0.59	-0.59	19.52	0.44
9	-0.47	614	-0.47	1214	0.0160	36	0.0160	36	-0.59	-0.59	19.69	0.15

The same data that comprise the first numerical example and are depicted in figure 3.15 have already been shown in figures 3.6 and 3.8 for the description of the PIT and MD area metrics. Analytically, the displacement and deformation measurements were taken at the middle of a simply supported beam that was loaded vertically with a load of 5kN (at the middle) and a temperature difference of 30°C. The spread in the data stems from the uncertainty of the input parameters; namely the Young's modulus and the coefficient of thermal expansion (CTE). The former was characterised by a Gaussian with a mean of 73.1 GPa and a standard deviation of 2.5GPa (CoV=3.4%) while the the CTE was characterised by a Gaussian centred on $22.8\mu\epsilon^{\circ}C^{-1}$ with a standard deviation of 1 $\mu\epsilon^{\circ}C^{-1}$ (CoV=4.4%). A total of 100 samples were taken from these inputs to represent the simulated responses and 10 samples were taken to represent the experimentally measured quantities.

The results of the comparison shown in figure 3.15 suggest that a nonzero value arises in both metrics even though both datasets stem from the same distribution. This could be considered the 2-D equivalent of the results of figure 3.10 and in a similar manner can be attributed to the lack of a sufficient amount of measurements. However, it is obvious that the transformed quantities have accurately captured the proximity between the two distributions resulting in low values.

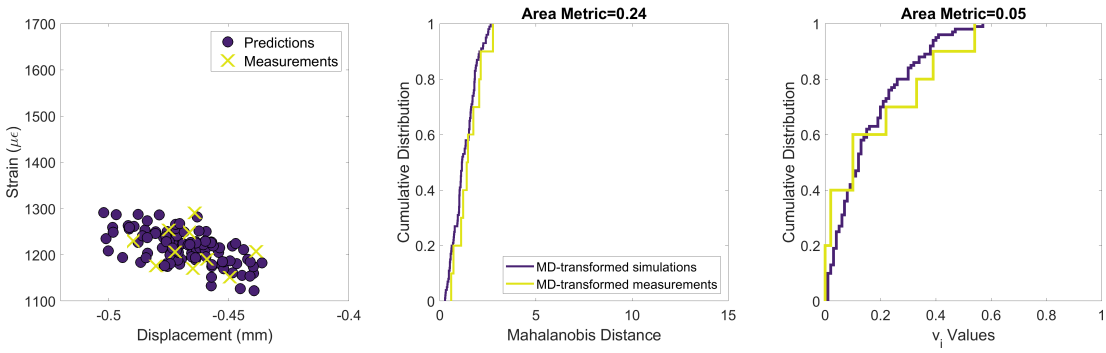

Figure 3.15: Example 1: The means, standard deviations and correlations are the same in both datasets. The simulated and measured responses are shown on the left. In the middle, the Mahalanobis-based transformation of the measurements and simulations is portrayed along with the result of their comparison on the title. On the right, the probability integral transformation of the simulations is shown in purple along with v-transformed measuremets.

Figure 3.16 depicts the effect of mismatching correlation coefficients on the metrics. In this case the correlation coefficient for the data generating function was zero for the measured quantities. The impact on the Mahalanobis distance area metric is minor, resulting in a value close to that of example 3 where the correlation coefficient was the same for both the simulated and the measured quantities. This similarity should be attributed to the limited number of measurements that do not suffice to properly establish the effect of mismatching correlations. On the other hand, a significant change is evident on the PIT area metric (from 0.30 in example 3 to 0.44 in example 4). This difference is attributed to the clustering of the measured quantities in example 4. Analytically, it can be seen in the same figure that the v-transformed measurements take values between 0.3 and 1, reflecting a limited coverage of the predictions in the x-axis (displacements). This is not the case for example 3 (figure. B.3) where the range of the measurements almost overlaps that of the predictions.

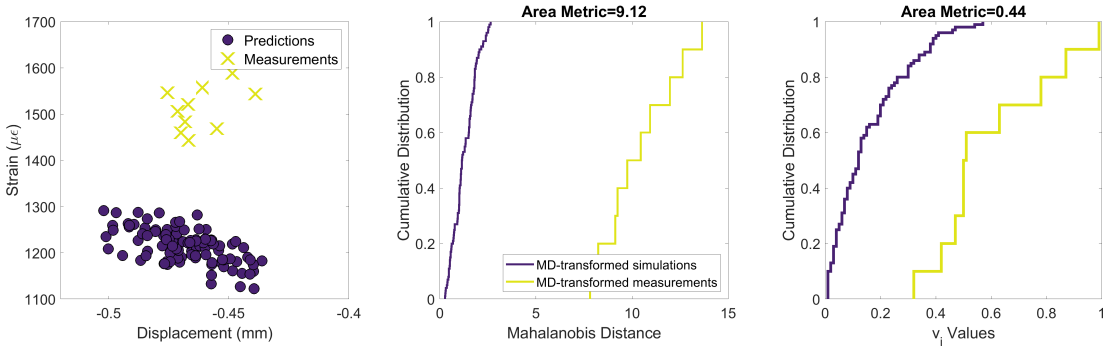

Figure 3.16: Example 4: There are differences in the means and the correlation coefficients between the datasets. In this case the correlation coefficient for the measured quantities is zero.

Finally, figure 3.17 shows the combined effect of bias in the measurements and strong negative correlation in the simulations ($\rho_{sim} = -0.95$). In this case two phenomena are simultaneously at play. It can be seen in the right graph that the empirical distribution of the v-values corresponding to the measured dataset is the indicator function in the domain $[0, 1]$. This is attributed to the fact that all the measurements are defined in a region where the predictions' distribution function is zero. This results in a cluster of zero-valued v-values and the vertical jump in their empirical distribution function at $v = 0$. The second phenomenon is associated with the shape of the predictions' PIT curve. Li et al. [60] misinterpreting the work of Genest and Rivest [61] suggest that in the extreme cases of absolute positive or negative correlation between two variables ($\rho = 1$ or $\rho = -1$) the resulting simulations' v-transformed distribution would be a uniform one and its distribution function will take the form of a straight, 45° line in the v-value space. Even though this is correct for the case where $\rho = 1$, it is wrong in the case of negative correlation where the simulations' v-transformed distribution function is the indicator function. What this practically means is that in the scenario where the simulation outputs are negatively correlated and there is also considerable amount of (downward-left) bias between the two datasets, as in this case, the resulting PIT area metric will be zero. This is an issue that Li et al. did not identify in their work and should be taken into account when using their technique.

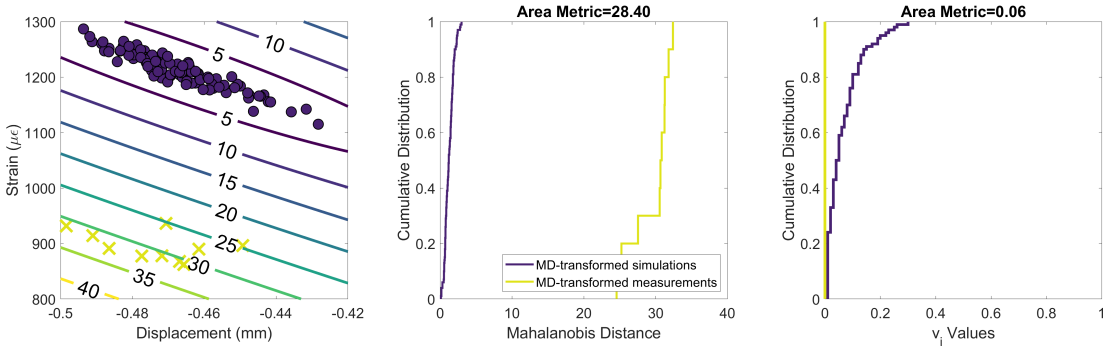

Figure 3.17: Example 7: There are differences in the means and the correlation coefficients between the two datasets. The correlation coefficient of the simulated dataset is -0.95 while the respective experimental one is -0.59.

Another limitation of the PIT area metric is dimensionality. Figure 3.18 depicts the multivariate PIT transformation of a Gaussian distribution ($\rho_{ij} = 0$) of increasing dimensionality. It can be seen that the distribution function of the v-values grows rapidly to one as the dimensionality of the problem increases thus limiting its capability to distinguish dissimilarities between distributions.

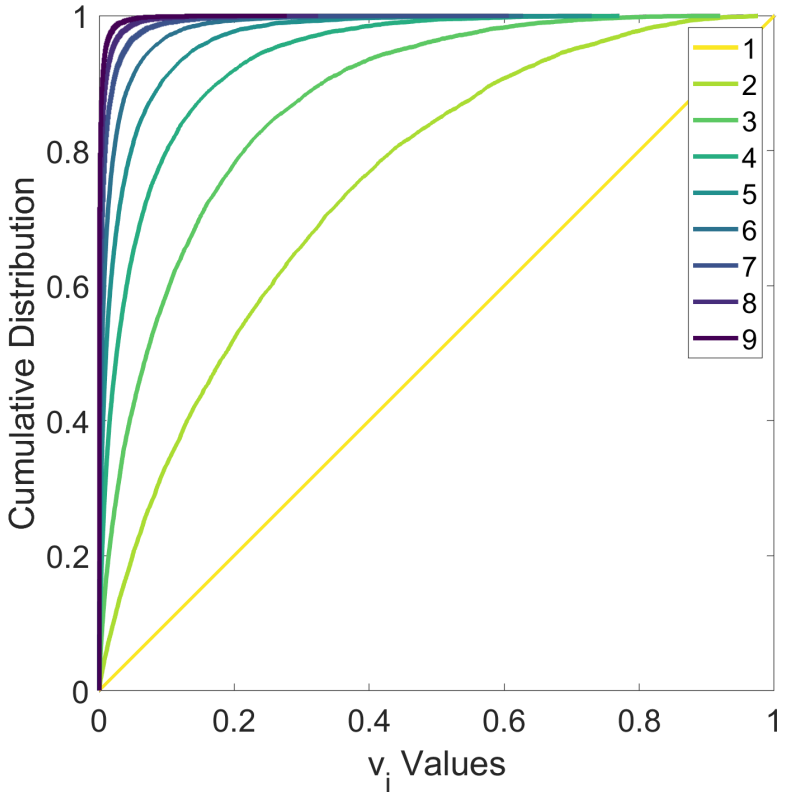

Figure 3.18: The PIT transformation of a Gaussian distribution of increasing dimensionality ($\rho_{ij} = 0$). The dimensionality of each curve is outlined in the legend

3.4 Hole-in-plate experiment

This application is on an aluminium plate with a hole in the middle. The plate was clamped on one end while a uniformly distributed load of 8 kN was applied on the other end. Strain measurements were simultaneously taken from six strain-gauges carefully positioned across its surface. The setup is demonstrated in figure 3.19 and represents a simple, yet powerful experiment where complex finite element models (FEM) or analytical solutions can be used for validation practice. The mechanical drawing of the geometry along with the strain-gauge numbering are shown in figure 3.20. A Finite Element Model was developed using the commercially-available software ABAQUS where a total of 18,500 triangular and quadrilateral shell elements (S3R,S4R) were used for the meshing. The uncertain parameters for this case were the Young's modulus, Poisson's ratio and the thickness of the specimen. A three-dimensional, uniform distribution whose parameters have been outlined in table 3.3 was used to characterise the above. Afterwards, one hundred samples were generated from the joint distribution using Latin Hypercube sampling [121] and were subsequently used as inputs in the computational model. The MATLAB code used for that process is by Minasny [125]. At the same time six sets of repeated measurements were taken using the same specimen. These represent the variation in the response of the structure and the embodied measurement uncertainty.

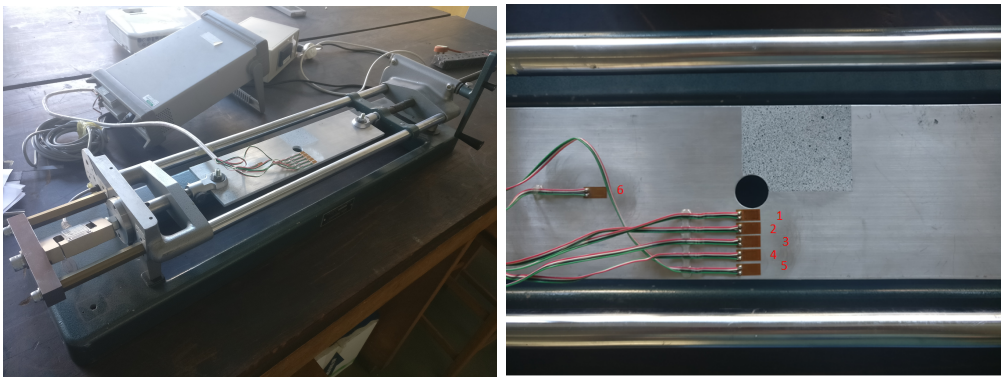

Figure 3.19: Hole-in-plate specimen setup (left) and strain-gauge numbering (right).

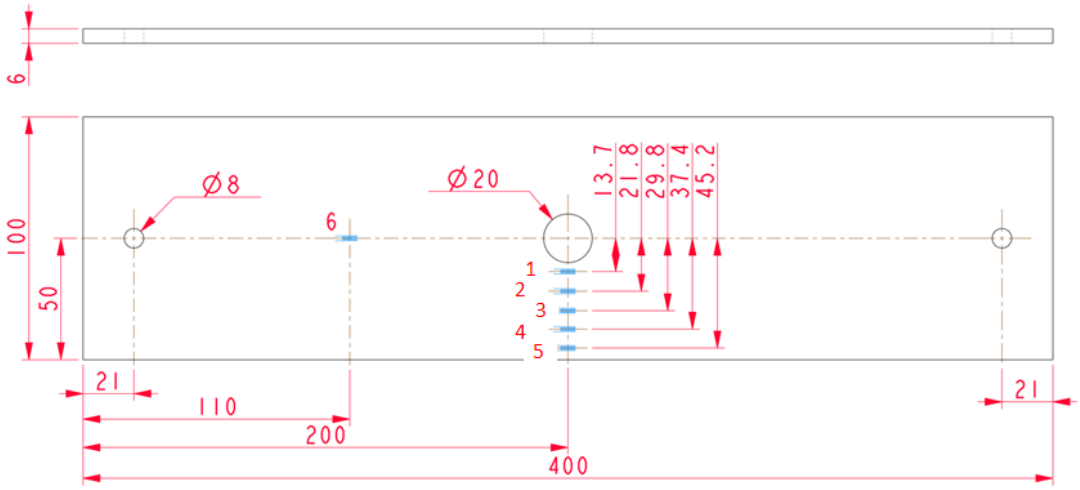

Figure 3.20: Hole-in-plate CAD drawing (dimensions are in mm).

Table 3.3: Hole-in-plate uncertain parameter characterization.

Parameter	Population distribution	Distribution parameters
$E(GPa)$:Young's modulus	Uniform $\sim U(a,b)$	$a = 60.0, b = 80.0$
ν:Poisson's ratio	Uniform $\sim U(a,b)$	$a = 0.2, b = 0.4$
$t(mm)$: Specimen thickness	Uniform $\sim U(a,b)$	$a = 5.9, b = 6.2$

Results

In order to achieve a straightforward model-experiment comparison the strain data from the model were averaged over the elements whose virtual location coincided with that of the strain-gauges. Figure 3.21 shows the results from the model and the experiment for each strain-gauge across 6 graphs. The text next to each experimental distribution function reports the strain-gauge identification (SG) according to figure 3.20. The experiment was repeated six times and each of those measurements for the respective strain gauge is represented by the appropriate measurement number (meas. no.). It can be seen that the predictions have a greater degree of variability compared to the measurements, a detail that can be ascertained from figure 3.22 where the readings of the first five strain gauges have been plotted against the distance from the hole edge. The mismatch between the predicted and measured variances is reflected by the error bars portraying a situation similar to the one depicted in figure 3.13. The high level of variability in the predictions can be attributed to the uncertainty in the model's inputs. Uniform distributions, whose parameters are described in table 3.3, were

selected to represent the epistemic uncertainty in the material properties as direct material characterisation information was missing.

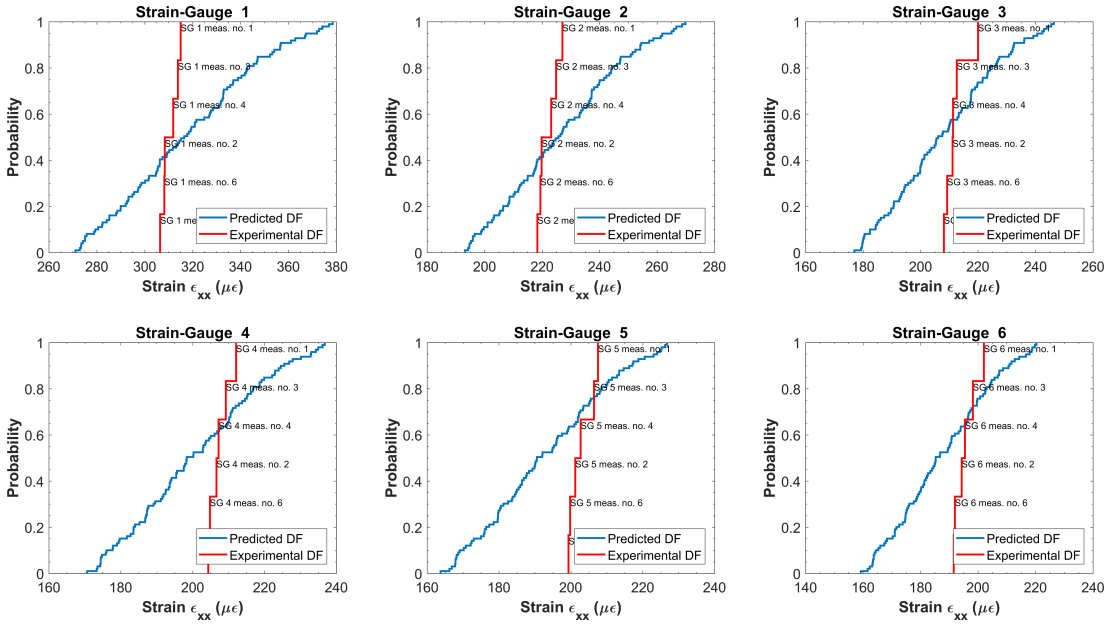

Figure 3.21: Distribution functions of the measurements against the predicted ones for the strain gauges depicted in figure 3.20.

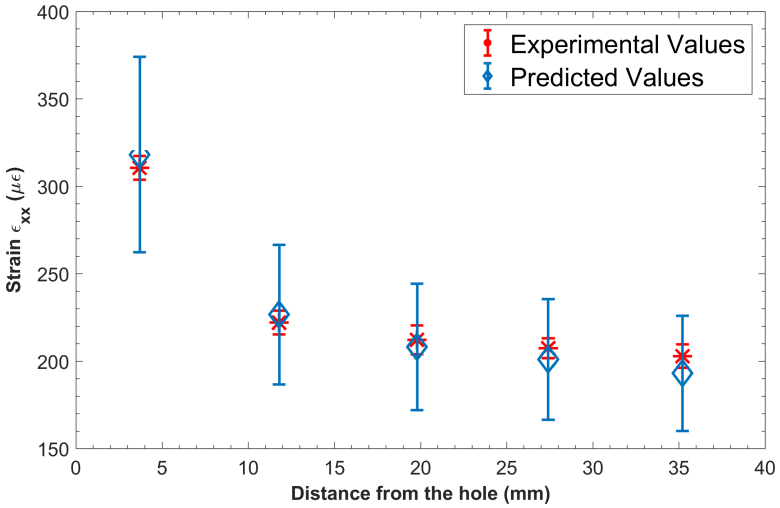

Figure 3.22: Model predictions against experimental measurements across the distance from the edge of the hole. The error bars show the range of values from the model and the experiment.

Even though a qualitative comparison seems straightforward in this case, the objective is to quantitatively assess the mismatch between measurements and predictions. To achieve that, the area metric and u-pooling were employed at each strain gauge location and the results of the comparison are available in table 3.4. Afterwards, the results across the strain gauges were pooled together

in an attempt to demonstrate how u-pooling can be used to assess a model's accuracy in multiple locations. This process is known as 'marginal' u-pooling as the measurements in each location are pooled together without considering potential correlations amongst the rest. The result of this process can be seen in figure 3.23. The limited spread of the measurements compared to the 45° curve is evident from the range of u-values in the abcissa. The u-transformed measurements lie between the 40^{th} and 78^{th} ($u = 0.40$ and $u = 0.78$ respectively) percentile of the model's predictions reflecting the data in figure 3.21.

Table 3.4: Hole-in-plate validation results.

strain-gauge	area metric ($\mu\epsilon$)	area metric 95% CI	u-pooling
1	21.6	[18.7,26.0]	0.23
2	14.6	[12.5,18.0]	0.22
3	13.2	[11.0, 16.3]	0.21
4	13.7	[11.8, 16.4]	0.23
5	13.9	[12.0, 17.4]	0.25
6	12.2	[10.6, 15.8]	0.22

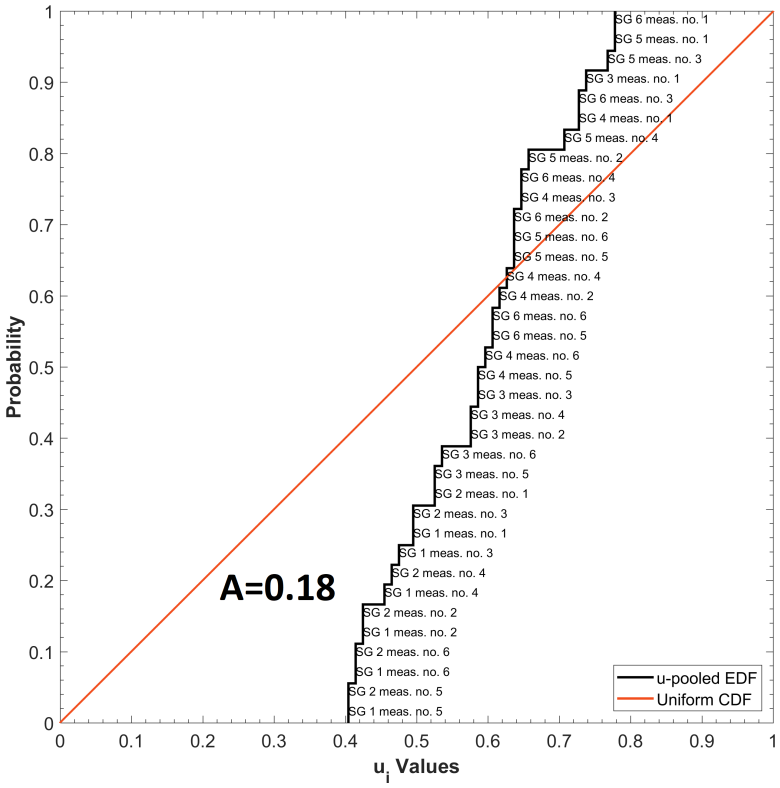

Figure 3.23: Marginal u-pooling of the strain-gauge measurements for the hole-in-plate experiment. The identifier corresponding to each strain-gauge measurement is shown next to it.

In order to demonstrate the effect of parameter uncertainty on model validation, the same model with updated parameters was run. This time the samples for the Young's modulus were taken from a uniform distribution that extended between 65 and 75 GPa. The results of marginal u-pooling across the strain-gauges are shown in figure 3.24 while the respective localised, gauge-based assessments can be seen in table 3.5. A visual comparison of figures 3.22 and 3.25 suffices to demonstrate that the updated parameters improve the model-experiment proximity. In addition to the visual comparison, the updated value of marginal u-pooling (0.16 against 0.18) acts as proof that the model with the updated parameters is a better representation of the real world. Comparing figures 3.23 and 3.24 it is obvious that the updated model predictions are more accurate, a fact reinforced by the spread in the measurements' u-values (ranging from the 28th to 97th percentile of the model's predictions). It is also evident that the measurements are right-tail biased compared to the predictions. This can be seen in figure 3.25 where for most of the strain-gauge locations, the experimental values seem to be lying on the uppermost region of the predicted range.

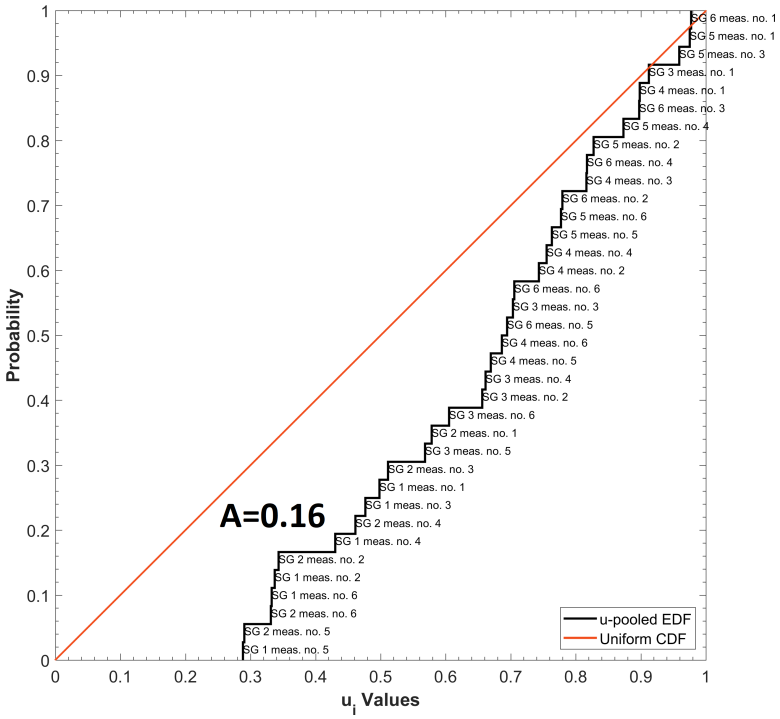

Figure 3.24: Marginal u-pooling across strain-gauge measurements for the updated model.

Table 3.5: Hole-in-plate updated model parameters validation results.

strain-gauge	area metric ($\mu\epsilon$)	area metric 95% CI	u-pooling
1	10.1	[9.2, 12.4]	0.20
2	6.4	[5.6, 8.7]	0.18
3	6.4	[5.8, 9.3]	0.21
4	8.2	[7.4, 10.6]	0.27
5	11.1	[9.2, 13.7]	0.36
6	9.1	[7.4, 12.1]	0.31

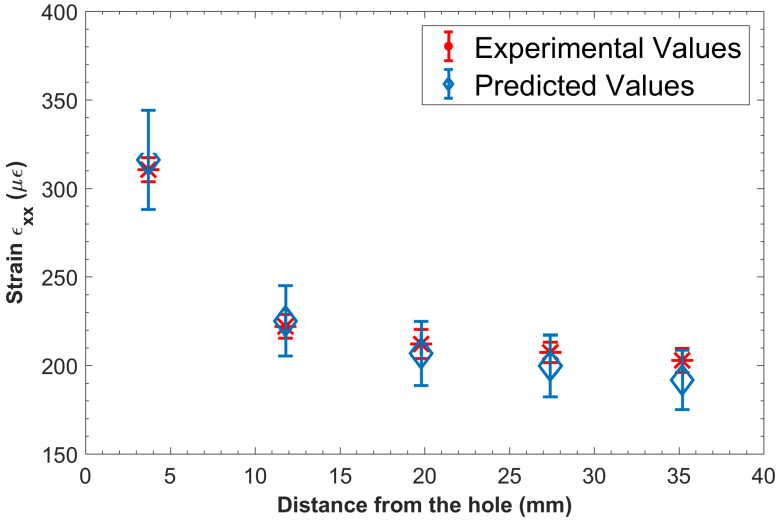

Figure 3.25: Model predictions against experimental measurements across the distance from the edge of the hole for the updated model. The error bars show the range of values from the model and the experiment.

In terms of local comparisons resulting from tables 3.4 and 3.5 it is evident that the area metric can capture the improvement stemming from the updated model parameters. On the other hand, the results from the use of u-pooling do not seem to follow the same pattern. Even though a reduction is evident on the first and second strain gauges, it seems that there is an increase in these values for the fourth, fifth and sixth strain gauges. These increases are attributed to the fact that u-pooling is more sensitive to differences in the means rather than differences in the variances between the two distributions. This local effect is cancelled out by strain gauges 1 and 2 when pooling them all together as in figure 3.24.

So far, during marginal u-pooling all the measurements were pooled together without considering the correlations among locations. However, it is known both theoretically and experimentally that the magnitude of strain, in the longitudinal

direction around a hole, is proportional to the distance from the edge of the hole. This phenomenon which is visible in figures 3.22 and 3.25 should be accounted for when making assessments of the model's accuracy. To establish this, each set of measurements across the 6 strain gauges was treated as a point in the 6-D space. Afterwards, the multivariate PIT and the Mahalanobis distance area metric were used to assess the model's accuracy while accounting for the emerging functional correlations. The result of the comparisons using these metrics can be seen in figure 3.26.

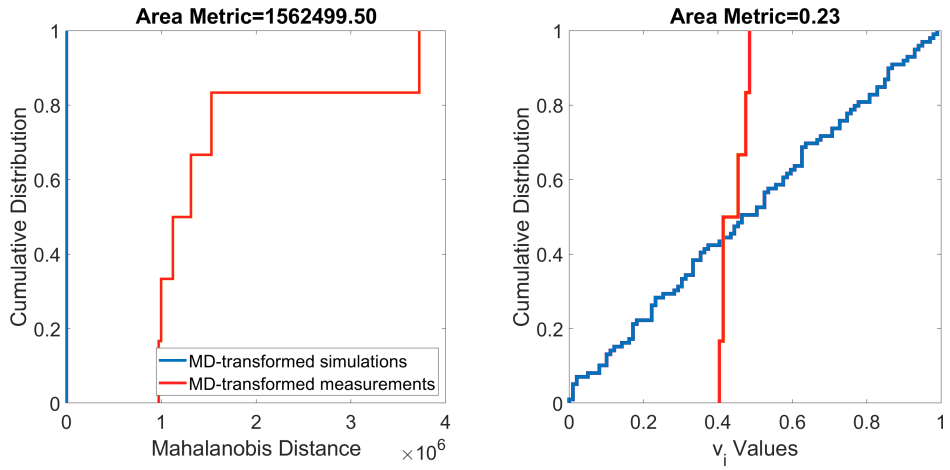

Figure 3.26: Mahalanobis-distance and probability interal transfrom area metric for the hole-in-plate experiment.

The left side of the figure demonstrates the Mahalanobis distance area metric. It can be seen that the MD-transformed measurements are in the range of millions. This is attributed to the fact that the simulations' covariance matrix is almost singular which is a result of the perfect linear dependence across the strain predictions in the strain gauge locations. For example when the $\epsilon_{\chi\chi}$ in strain gauge 1 increases so do the corresponding ones in the rest of the strain gauges as it can be seen in figure 3.27. The singularity in the covariance matrix penalises measurements that may even slightly deviate from this pattern (for example due to the existence of measurement error or minor changes in the boundary conditions). This outcome is similar to the one shown in figure 3.17. On the other hand, it can be seen in the right side of the same figure that the PIT-transformed simulations follow a 45°curve. This distinct pattern stems from the perfect positive linear correlations emerging across the simulated strain-gauge outputs. The measurements' v-values are clustered near 0.5 resulting in a PIT area metric of

0.23. This outcome qualitatively resembles the one of figure 3.23.

The presence of collinearities, as shown in figure 3.23, can be assessed by the rank of the covariance matrix. The rank of a matrix, in general, provides information about the dimension of the space spanned by its columns (and by its rows equivalently). This can act as an indicator of the linear dependency across its variables. In this case, the rank of the covariance matrix is 5 denoting it is rank-deficient given that its full rank would be 6 if the variables were linearly independent. This finding is reinforced by the condition number of the matrix whose large value ($3.09 * 10^{15}$) also points to singularity.

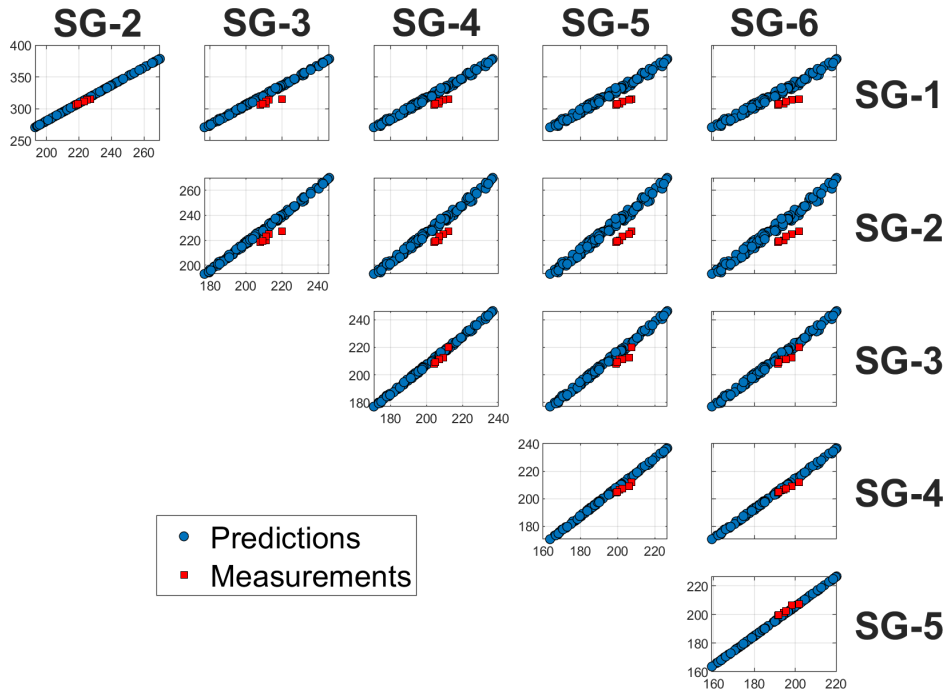

Figure 3.27: Monte Carlo simulation outputs plotted against the measurements for the designated strain-gauge identifications. The units are $\mu\epsilon$.

3.5 I-beam

In this section the probabilistic model validation approach described earlier is applied to full-field data. An important component of this process is the extraction of features from the measured and the predicted data to enable the comparison of datasets which are defined in different grids. In this case, this is achieved via the decomposition of the field-data with the aid of 2-D Chebyshev shape descriptors (SDs) into a vector of coefficients, the magnitude of which reflects their contribu-

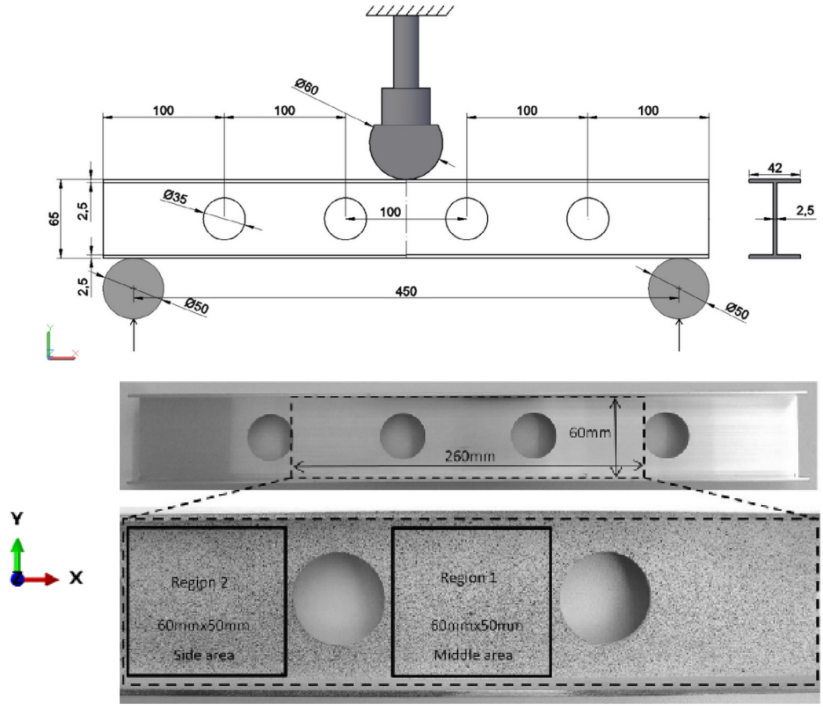

Figure 3.28: I-beam geometry (dimensions in mm) and loading (top). The area (Region 1) defined by the rectangle in the middle of the beam was used as the region of interest (bottom) [9].

tion in the reconstruction of the dataset. The resulting feature vectors from the predictions and the measurements are capable of representing the underlying data without significant information loss. In this section a feature-based assessment will be demonstrated, initially with the aid of marginal u-pooling and afterwards through the PIT and MD area metrics.

It is important to state that only synthetic data have been used in this example. The specimen used is an aluminium beam with an I-shaped cross-section and is loaded in three-point bending. The beam rests on two cylindrical rods of fifty-mm diameter that move upwards against a fixed, sixty-mm diameter rod on the top of the beam. The drawing of the specimen along with the loading is shown in figure 3.28. The selected region of interest (ROI) where the assessment will take place is a rectangular 60x50 mm area in the middle of the beam. The methodology will be demonstrated using displacements in the y-direction.

A total of one hundred simulations were run; each one reflecting the uncertainty in the Young's modulus and the Poisson's ratio of the beam's material. They were both characterized using uniform distributions and their parameters can be seen in table 3.6. As stated earlier only synthetic data were used for

this case; for that reason, five simulation results were used to represent the 'experiment' while the rest (ninety-five) represented the simulation outcomes. The decomposition of the displacement data took place afterwards. Four hundred (400) shape descriptors were used for the decomposition of the displacement data initially, but it was found that only nine of them were sufficient for their reconstruction without any significant information loss. To determine the number of shape descriptors that can adequately represent the data in the feature vector space, the recommendations made by the CEN guide [8] for the validation of computational solid mechanics models were implemented. It is suggested that the quality of the reconstruction of a data field, based on the features extracted using a decomposition technique, should be assessed using the average squared residual and that the average residual should not be greater than measurement uncertainty, u_{meas} obtained from a calibration of the measurement system. The resulting displacement fields can be seen in figure 3.29.

Table 3.6: I-beam uncertain parameter characterization.

Parameter	Population Distribution	Distribution Parameters
$E(GPa)$:Young's modulus	Uniform $\sim U(a,b)$	$a = 60.0, b = 80.0$
ν:Poisson's ratio	Uniform $\sim U(a,b)$	$a = 0.2, b = 0.4$

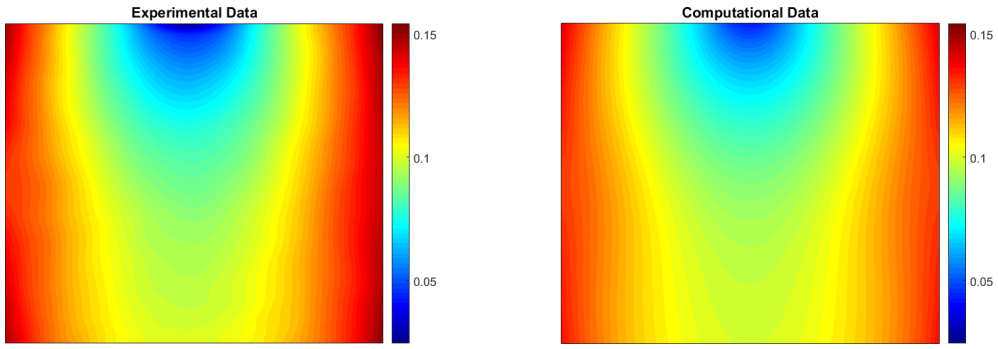

Figure 3.29: I-beam y-displacement measured and predicted fields.

Two cases were employed to assess the capability of the metrics to accurately distinguish measurements with similar or greater level of variability. For the first case, the five synthetic 'experimental' datasets were used without any modification. For the second case, the synthetic data were built according equation 3.16. This means that the experimental data for the second case are expected to have

the same mean value as the simulation outcomes and their standard deviation should be five times the simulations' standard deviation.

$$C_{E_{j,\lambda}} = \bar{x}_{S_\lambda} + 5S_{s_\lambda}X,\ X \sim \mathcal{N}(0,1) \tag{3.16}$$

where:

$$\bar{x}_{S_\lambda} = \frac{\sum_{k=6}^{100} C_{S_{k,\lambda}}}{95}, \qquad S_{s_\lambda} = \sqrt{\frac{\sum_{k=6}^{100}(C_{S_{k,\lambda}} - \bar{x}_{S_\lambda})^2}{94}} \tag{3.17}$$

λ represents the λ_{th} Chebyshev Coefficient, $C_{E_{j,\lambda}}$ represents the j_{th} 'experimental' result and the λ_{th} Chebyshev coefficient. $C_{S_{k,\lambda}}$ represents the k_{th} 'simulation' output and the λ_{th} Chebyshev coefficient. $\mathcal{N}(0,1)$ represents a sample from a normal distribution ($\mu = 0, \sigma = 1$).

Results: first case

Figure 3.30 shows the results from the first case. The bar height represents the average of each coefficient for the experimental measurements and the Monte Carlo simulations, while the green error bars represent the range of values respectively. As stated earlier, only nine coefficients whose identifiers are shown in the x-axis were used for the reconstruction of the data fields. To get a better idea of the mismatch between measurements and predictions figure 3.31 depicts the distribution functions of the largest three shape descriptors. Even though all of them originate from the same distribution, it is clear that the step-wise experimental distribution functions deviate from the simulation distribution functions. This phenomenon is similar to the one of figure 3.10. Pooling the coefficients together results in figure 3.32. It can be seen that all of the experimental results lie within the simulations' predicted range. Moreover, the u-pooled value (0.10) shows that there is a good level of agreement between the simulation and the experiment.

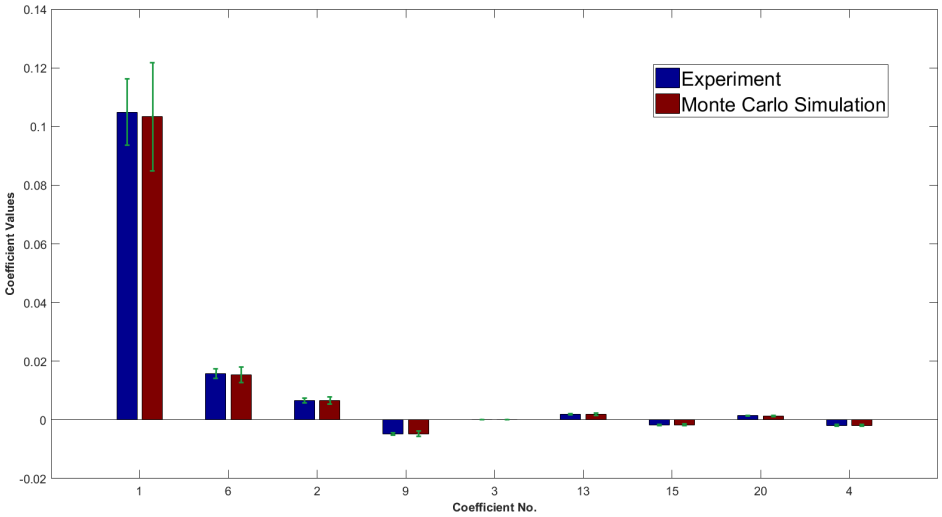

Figure 3.30: I-beam case 1: bar chart depicting the Chebyshev coefficients corresponding to the measured and predicted fields respectively. The height of each bar corresponds to the average while the error bars reflect the range in each coefficient.

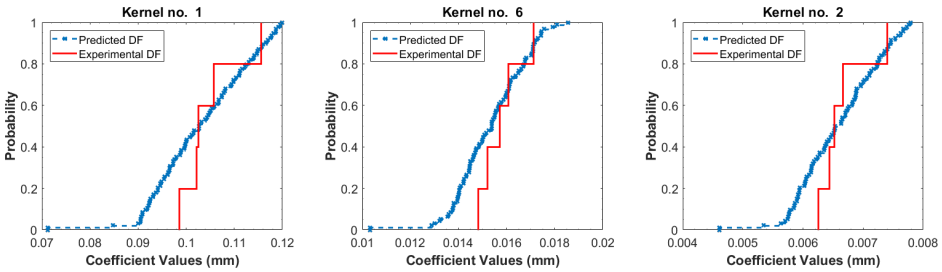

Figure 3.31: I-beam case 1. Measurements against predictions for the three largest Chebyshev coefficients depicted as distribution functions.

However, this calculation is based on the marginal representation of the data. This effectively means that each experimental shape descriptor was pooled with respect to its corresponding simulated one independent of the rest. Even though this form of data pooling can provide an overview of the shape descriptors that are in agreement, it may lead to wrong conclusions. To account for this issue and to make up for potential correlations between shape descriptors the pooling should be done jointly. In this case this was established with the aid of the Mahalanobis distance and PIT area metrics; the former resulting in a value of 0.32 and the latter in a value of 0. The value of the MD area metric can be explained by the fact that both datasets come from the same distribution and can thus be interpreted as an indication of the model's capacity to represent the real world, especially when considering the wide range of values corresponding

to the simulation outputs and the measurements (shown in the left side of figure 3.36). On the other hand, the (false) value of the PIT area metric is attributed to the fact that negative correlations emerged across the shape descriptors. The incapacity of the PIT area metric to accurately quantify the distance between two multivariate distributions in similar cases, an issue illustrated also in figure 3.39, means that its use in probabilistic model validation should be avoided.

From a practical standpoint, it should be highlighted that the various metrics provide a quantitative evaluation of the capacity of the models to represent the real world and should not be regarded as the ultimate basis for decision making. It shall be stated that different problems, depending on their significance, require different thresholds or acceptance criteria and the use of the corresponding metric should be adjusted accordingly. For example, the MD value stated above could be considered acceptable for a low-consequence effect in case of failure, while it could prove to be catastrophic in a case of failure of a primary load-bearing component in an aircraft. Moreover, the behaviour of each metric is subject to problem-specific details, making it apparent that the selection of these thresholds or accuracy requirements should be treated in a case-by-case approach.

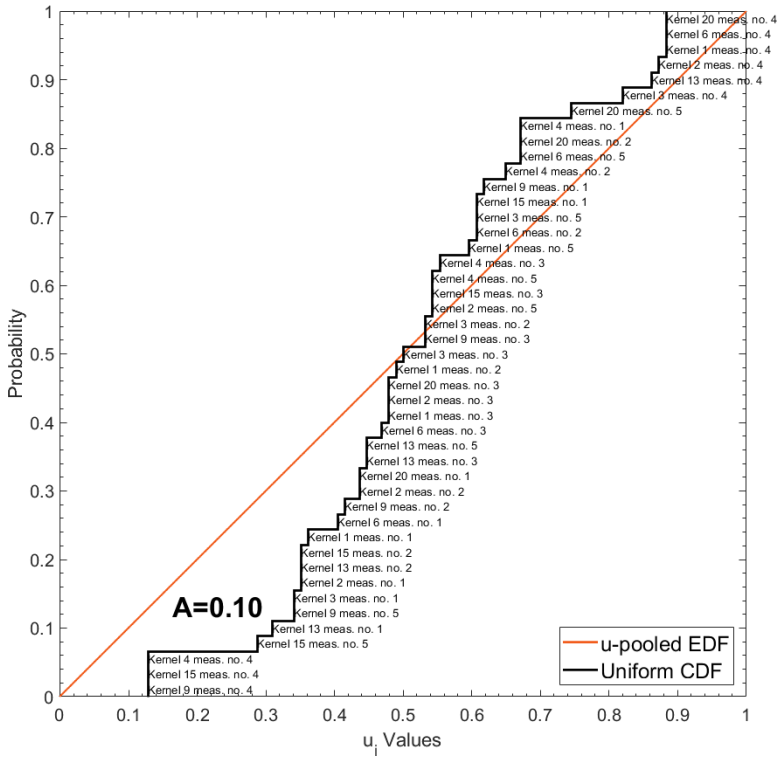

Figure 3.32: I-beam case 1: marginal u-pooling. The shape descriptors corresponding to the measurements have been marginally u-transformed by the respective predictions.

Results: second case

Respectively figures 3.33 and 3.34 show the results for the second case. It is obvious from both figures that the variability in the experimental shape descriptors is larger than the one in the predicted shape descriptors. This observation is enhanced in figure 3.35 where all the coefficients have been marginally pooled together. The shape of the u-transformed simulations reflects that only a few (12/45) of the measured shape descriptors are within the range of the corresponding predicted ones. The rest have values that are either larger than the range of the predicted ones, and their u-values are one, or have values lower than the smallest predicted ones and their u-values are zero. The corresponding u-transformed shape descriptor identifiers can be seen in the same figure. The information in this figure could be used as an indicator of shape similarity for applications where structural components are involved. For example, in modal analyses of components with simple geometries, where a small number of Chebyshev descriptors is needed to represent the underlying mode shapes [123], this

figure could inform engineers whether the measured modes are within the range of the predictions, providing information-rich insight in a simple graph.

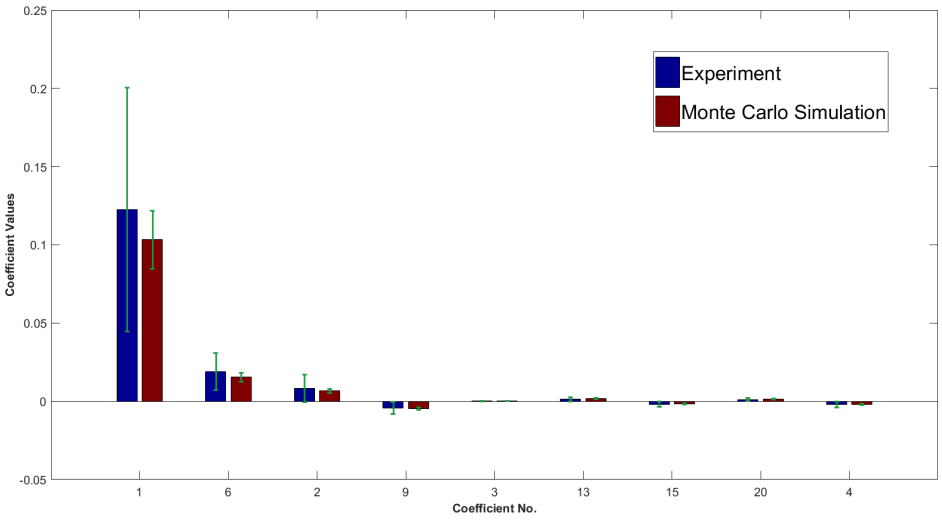

Figure 3.33: I-beam case 2: bar chart depicting the Chebyshev coefficients corresponding to the measured and predicted fields respectively. The height of each bar corresponds to the average while the error bars reflect the range in each coefficient.

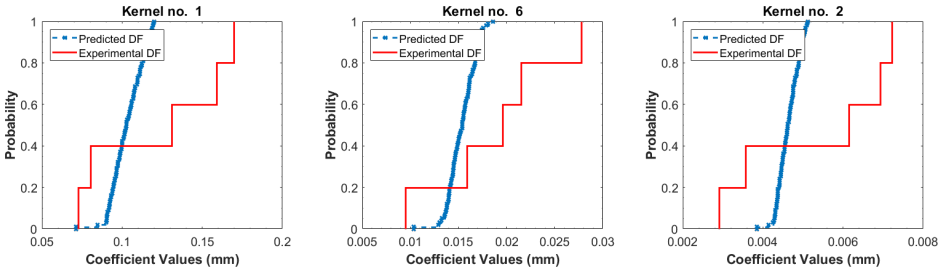

Figure 3.34: I-beam case 2: measurements against predictions for the three largest Chebyshev coefficients depicted as distribution functions.

The MD and PIT area metrics were calculated for this case as well; the former having a value of 13660 while the latter, similar to the previous case having a value of 0. The value of the MD area metric is attributed to the strong correlations emerging across the predictions' shape descriptors, in a manner similar to the numerical example of figure 3.17. This means that a measurement located away from the predictions and the hyper-ellispsoids formed by their Mahalanobis distance loci (such as the ones shown in figure 3.8, will be highly penalised, resulting in the large values shown in figure 3.36.

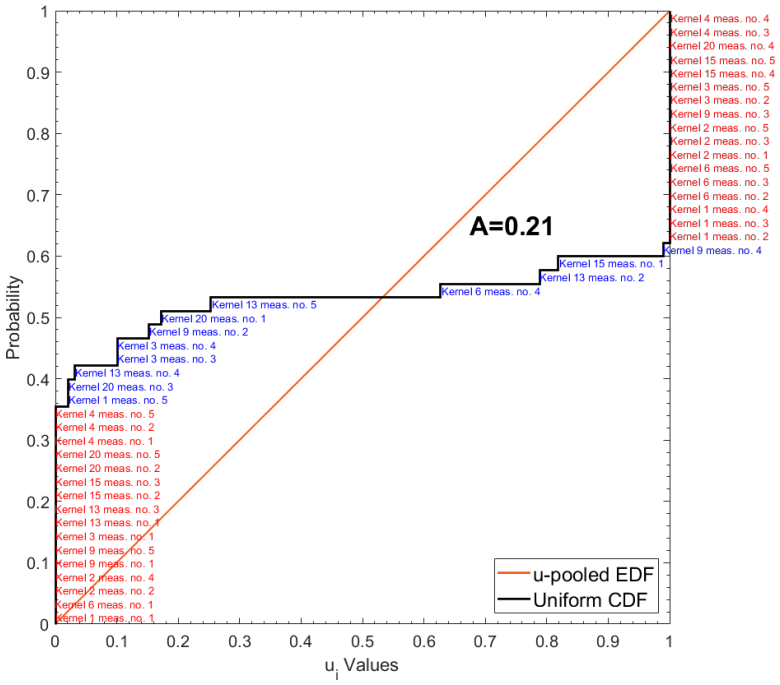

Figure 3.35: I-beam case 2: marginal u-pooling. The shape descriptors corresponding to the measurements have been marginally u-transformed by the respective predictions. The red and blue-coloured text correspond to measured shape descriptors outside and inside the predictions' range respectively.

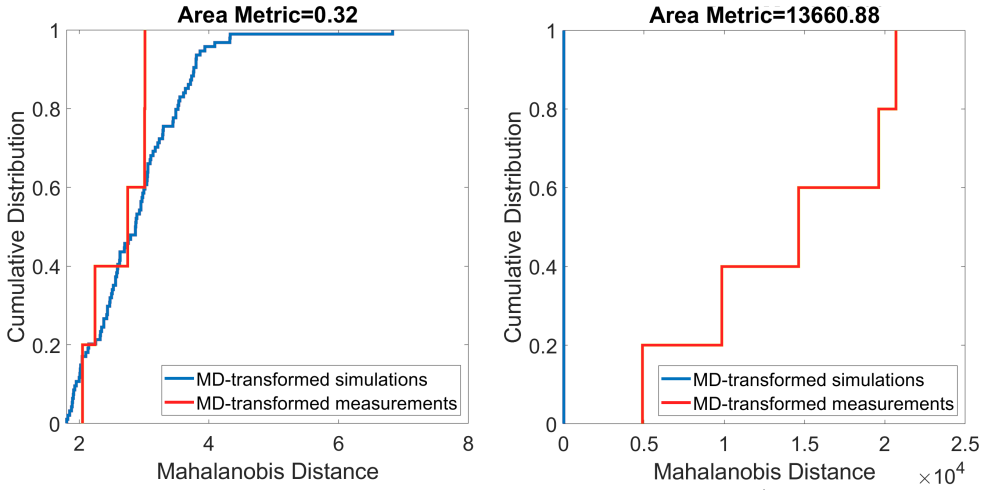

Figure 3.36: Mahalanobis distance area metric for cases 1 (left) and 2 (right) of the I-beam dataset.

3.6 Discussion

The analysis outlined in this chapter aimed to assess metrics for quantifying the proximity of probabilistic model outputs against experimental measurements. Starting with point measurements at a single location and leading to full-field

measurements across a region, the analysis can be described from the perspective of comparing two distributions: one corresponding to the experimental measurements and one corresponding to the outputs of the probabilistic model. To quantitatively assess the discrepancy between the two, several metrics were assessed. These metrics, depending on the form of the measured quantities and the requirements of validation, can be distinguished in univariate and multivariate. A number of numerical and engineering examples were used to demonstrate the potential and limitations of four metrics: the area metric, u-pooling, Mahalanobis distance and probability integral transform area metrics.

3.6.1 Behaviour of the area metric and u-pooling

Starting with the univariate metrics (area metric and u-pooling) at a single validation site, it was shown that a nonzero value surfaced during the comparisons even when the two distributions were the same and the number of the measurements was high (e.g. $N_{exp} = 1000$) as shown in the top rows of table 3.1. In the same table it is obvious that the value of the area metric is influenced slightly more by differences in the means of the two distributions rather than differences in their standard deviations. For instance, the area metric of example 9 ($|\Delta\mu_{(exp-sim)}| = 3\ \mu\epsilon$) is 3.27 while the corresponding value for example 3 ($|\Delta\sigma_{(exp-sim)}| = 3\ \mu\epsilon$) is 2.75. The trend is repeated in examples 5 and 13 and is reinforced by figure 3.14 where the slope of squares that correspond to differences in the means is greater than the slope of rhombi that correspond to differences in the standard deviations.

On the other hand, the behaviour of u-pooling under different parameter combinations is more complex. In the same figure it can be seen that when the differences across the means of the two distributions is zero ($\delta\mu_{exp-sim} = 0$ - purple rhombi) the value of u-pooling is capped to 0.25, while an almost random dispersion of u-pooling values surfaces when the differences across the standard deviations are zero ($\delta\sigma_{exp-sim} = 0$) . To improve the understanding of this situation, figure 3.37 has been plotted using the data of figure 3.14. The left graph demonstrates the effect of means (experimental and simulations) on the value of u-pooling (reflected by the colouring of the data) when $\delta\sigma_{exp-sim} = 0$. It

can be seen that a pattern of increasing u-pooling values emerges when moving in a direction perpendicular to the main diagonal. As expected, when the distance between two Gaussian distributions increases, given that they have the same standard deviations, so does the u-pooling value from their comparison. In the right side of the figure a smoother pattern emerges as ones move away from the main diagonal. This demonstrates that the value of u-pooling is less sensitive to differences in the standard deviations of the two distributions when their means are same. Moreover, the u-values for this case are bounded to 0.25 which means that even when the differences in their standard deviations are as high as 50 (bottom right corner) the effect on the outcome of u-pooling is limited.

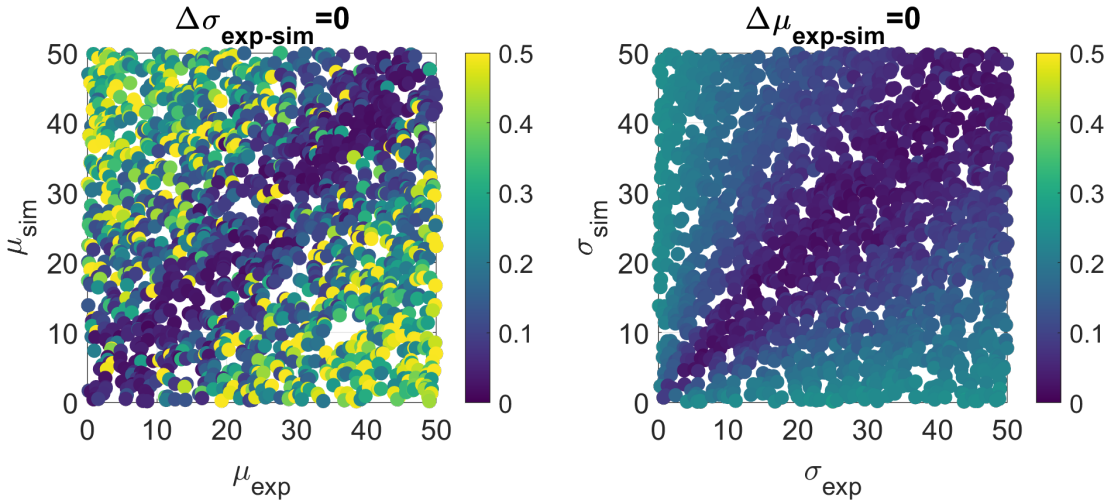

Figure 3.37: The effect of the distributions' means on the value of u-pooling when $\Delta\sigma_{exp-sim} = 0$ is shown on the left. On the right side the effect of the distributions' standard deviations when $\Delta\mu_{exp-sim} = 0$ is shown.

Expanding the analysis in the region where $\Delta\sigma_{exp-sim} = 25$ and $\Delta\mu_{exp-sim} = 25$ results in a situation change. The results shown in the left side of figure 3.38 depict a repetition of a weaker form of the previous pattern. The slow change in values perpendicular to the main diagonal should be attributed to the fact that the two distributions almost certainly overlap (given that $\Delta\sigma_{exp-sim} = 25$ and the range of means is $[0, 50]$). On the right side it is obvious that the previous pattern has changed; small values of σ_{exp} or σ_{sim} lead to large u-pooling values. This is natural as distributions that are distanced apart ($\Delta\mu = 25$) and are highly concentrated (low σ values at the bottom left corner) will have zero overlap resulting in high u-pooling values. The trend of decreasing u-pooling values

continues diagonally as the variance in both distributions increases for a constant $\Delta\mu_{exp-sim}$ value.

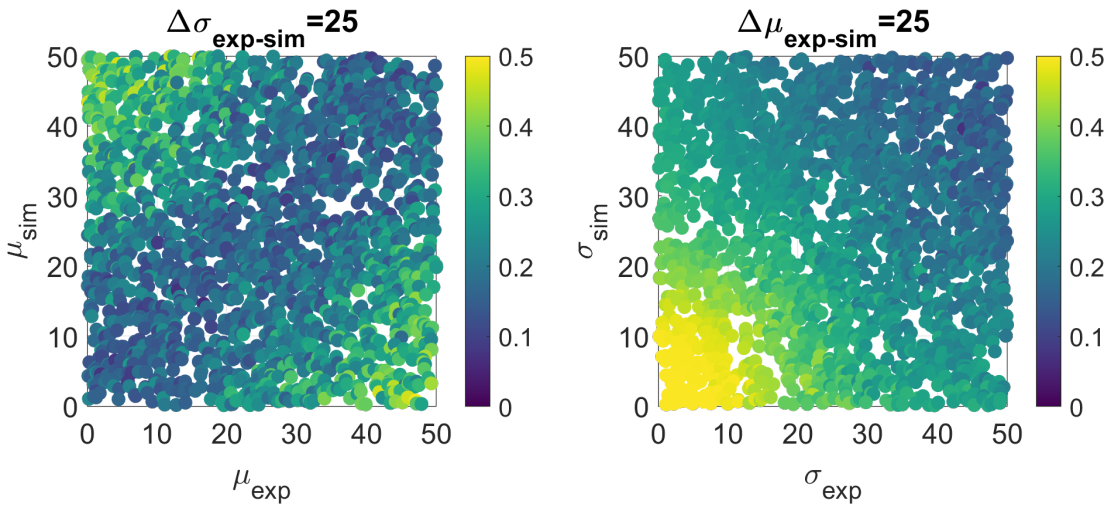

Figure 3.38: Similar to figure 3.37 when $\Delta\sigma_{exp-sim} = 25$ and $\Delta\mu_{exp-sim} = 25$

3.6.2 Multivariate validation and the effect of correlations among variables

Even though most of the validation processes are performed using a single output at a validation site, there may be cases where a model's predictions should be assessed against multiple types of measurements (such as temperature and acceleration) or against measurements across spatial locations. In these cases, it would be inappropriate to assume that each outcome is independent from the rest. To account for potential dependencies between variables or across spatial locations a multivariate metric is needed; in the examples demonstrated this was achieved (partially) using the PIT and MD area metrics. It was found that even though the PIT area metric can provide accurate assessments in multiple occasions, there are many cases where its shortfalls outweigh its benefits. Analytically, when strong negative correlations, such as the ones shown in figure 3.17 (example 7) and figure 3.39 emerge, the outcome can be severely flawed. Moreover, the negative effect of dimensionality on the results of the PIT metric was witnessed in figure 3.18. Analytically, the sensitivity of the PIT metric to identify deviations between multivariate distributions is greatly reduced when the number of dimensions is greater than two. Even then, the interpretation of the result is only subjective as

demonstrated in a similar manner by its univariate analogue (u-pooling) earlier.

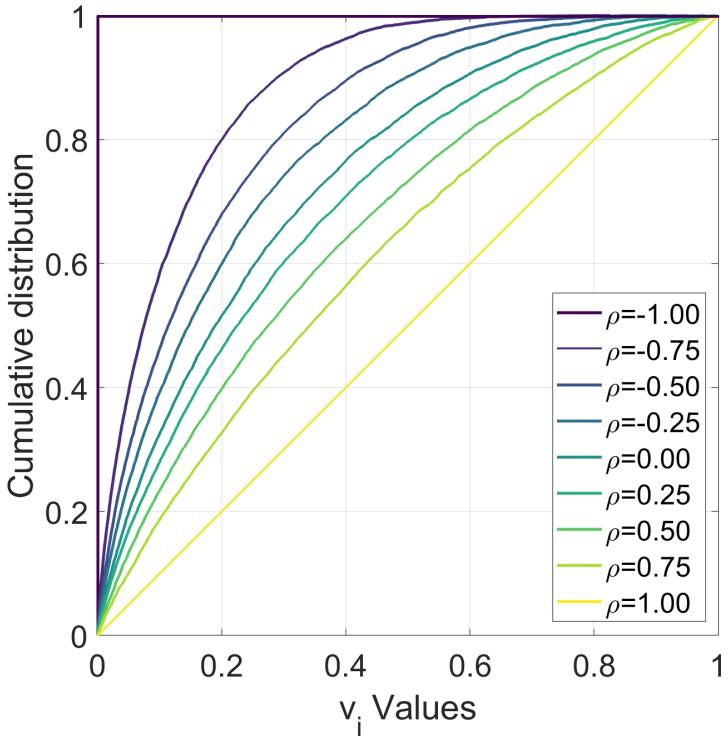

Figure 3.39: The effect of correlations on the probability integral transformation of a 2-D Gaussian distribution.

On the other hand, the use of marginal u-pooling demonstrated in figures 3.23 and 3.35 where each measurement is pooled independent of the rest can be exploited to aid decision makers identify which measurements deviate from the predicted ranges. However, this practice should be performed with great care, especially when communicating its results. In cases where the dimensionality of the data is reduced, for example through a decomposition technique (e.g. orthogonal decomposition using Chebyshev shape descriptors or principal component analysis), the coefficients representing the initial data in the new reduced-dimensionality space should be jointly assessed. To better demonstrate this issue, consider a case where all the experimental shape descriptors, except the first (corresponding to the mean of the underlying data), are in the range of the respective predicted ones. Pooling all the shape descriptors marginally would result in a low value, meaning that the two datasets are similar. However, this similarity could prove to be false, due to a potentially large difference in the means (first shape descriptor). A similar case could emerge in the case of principal component analysis where a significant deviation in the first few components, between the meas-

urements and predictions, if undetected, could lead to misjudgements regarding the capability of the model to represent the real world. These issues emerge from the fact that the different features carry different amounts of information and pooling them marginally together would nullify their contribution.

The limitations identified in the PIT area metric were alleviated with the Mahalanobis distance area metric as shown across the examples. A drawback of the technique, shown in figure 3.36, is the unintuitive transformation of experimental measurements with respect to the simulations when the outputs of the latter are highly correlated. A suggestion to address this issue would be to reduce the number of variables against which the assessment is performed. This stems from the fact that highly correlated variables represent information redundancy, an example of which can be found in the hole-in-plate experiment. In that case the simulation output in a location was perfectly correlated to the outputs of the rest as shown in figure 3.27. This could mean that a single strain gauge combined with the correlation structure across the rest of the strain gauges would suffice to assess the quality of the predictions.

3.7 Conclusions

This chapter reviewed some of the available techniques for probabilistic model validation. Compared to traditional forms of validation where a deterministic model output is matched against a measurement, these techniques allow engineers to account for the various sources of aleatory uncertainty in their models and to provide quantitative assessments of their capability to represent the real world. The four techniques: area metric, u-pooling, MD and PIT area metrics were assessed across a series of numerical and engineering examples. The conclusions drawn from this chapter could set the basis for the selection of a metric to assess univariate or multivariate probabilistic model predictions.

For the case of univariate data it was found that both the area metric and u-pooling can accurately assess the similarity of two univariate distributions; the former being equally sensitive to differences in the means and standard deviations when comparing two Gaussian distributions while the latter followed a complic-

ated behaviour. Its results should be thoroughly reviewed before communicated for decision making.

U-pooling in its marginal form was also used to aggregate measurements across different locations and shape descriptors across multiple measurements. Even though this practice should be avoided when a quantitative assessment is sought, it can be used as a screening method for identifying measurements that are within the predictions' range.

For the case of model validation using multivariate data (e.g. multiple response outputs) it was found that the capability of the PIT area metric to provide correct assessments of a model's predictions is greatly diminished as the dimensionality of the problem (number of response quantities) increases. This issue is aggravated in certain cases (e.g. response quantities are negatively correlated) thus leading to false results. On the other hand, the MD area metric can cope with higher dimensions while accounting for correlations among the response quantities.

In addition to point measurements, the applicability of the existing validation metrics was extended to full-field measurements with the aid of orthogonal decomposition using Chebyshev polynomials. The results from the application of the MD and PIT area metrics to synthetically generated data suggest that the MD area metric can be effectively used to pool multiple measurements and predictions, each representing a point in the multivariate space, and assess their proximity. The outcome of this metric can be used to inform the selection of the best among competing models.

Finally, it should be mentioned that only the effect of aleatory uncertainty was considered when assessing probabilistic models in this chapter. The effect of measurement error, inherent in any form of measurement, and its accurate representation in the feature vector space for model validation, will be analysed in the upcoming chapters.

Chapter 4

Transformation of measurement uncertainties into feature vector space

---

4.1 Introduction

As described in the literature review, a powerful method to determine the degree of similarity across spatial datasets is by decomposing them into a number of coefficients, each of which corresponds to a specific shape descriptor, thus assembling a feature vector. These shape descriptors may be characterised by a set of polynomials, e.g. Zernike, Krawtchouk, Legendre, Chebyshev, or in the case of principal component analysis (PCA) are resulting from the data itself in an optimal manner. Even though various techniques have been introduced to characterize the uncertainty rising from a small number of samples in the coefficients of PCA, such as by [126] and [127], an unresolved issue is the representation of the measurement uncertainty in a corresponding low-dimensional form.

No measurement is exact and the uncertainty in measurements can influence decisions about the reliability of simulations, the safety of processes, the quality of manufactured components or affect policy making; thus, leading to socioeconomic consequences. The challenge addressed in this chapter is the representation of measurement uncertainty, which can be constant or spatially varying, in the low-

dimensional form often used to extract features from information-rich data fields. The proposed method involves no assumptions about the probability distribution of the measurement uncertainty and is independent of the feature extraction process as demonstrated in three examples.

Overall, the process can be viewed as a function whose inputs are the spatial field of measurements, their uncertainty and the decomposition to the low-dimensional space; while the output is a distribution representing the measurement uncertainty in the low-dimensional space. The distribution is obtained via a chain of comparisons of the spatial measurement data with synthetically generated datasets, as will be described in the next section. While, in the subsequent section, the application of the new method to three examples of increasing complexity is presented. The first example consists of fields of displacements in an aluminium beam subject to three-point bending, measured using a digital image correlation system [9] with a measurement uncertainty that was spatially constant. This relatively straightforward example allows an in-depth explanation of the method and a graphical representation of the results using a simple displacement field and then, using a more complicated displacement field, a comparison with previously established recommendations for the validation of computational solid mechanics models. The second example is more complicated with spatially varying measurement uncertainties associated with soil-moisture measurements resulting from a Kriging analysis of sparse measurement stations at the Heihe River Basin in China [128]. The final example introduces two additional factors in the uncertainty: gaps in the data and a progressive reduction in uncertainty over time as the measurement acquisition technology is improved. It involves global oceanographic temperature fields obtained monthly over eleven years from 2002 to 2012 [129],[130].

4.2 Methodology

4.2.1 Transformation of the measurement uncertainty

The overall goal is to enable the quantitative comparison of information-rich data sets in order to inform rational decisions with consequences. In many cases, this

will involve comparing fields of measurement data with either other sets of measurement data or predictions from a model; and the decision will be influenced by whether or not the difference between the data fields is significant, which requires knowledge of the associated uncertainty in the data. When the comparison is performed by decomposing the data to a low-dimensional form, then it is necessary to make a quantitative assessment of the difference between the corresponding components in the low dimensional space, which requires transforming the measurement error into the same space. However, in practice this transformation of the uncertainty is not performed due to a gap in knowledge about an appropriate methodology. A new methodology that uses approximate Bayesian computation [31] and results in a distribution representing the measurement and its uncertainty in the component or feature vector space will be described.

The overall methodology for transforming the spatial data and its uncertainty to its low-dimensional form involves the steps shown in figure 4.1 and the approximate Bayesian computation in figure 4.2. Initially in figure 4.1, the dataset is decomposed to represent the data field in a lower dimensional form as a feature vector or set of components. The methodology is independent of the mathematical transformation or decomposition used in this initial stage; and, this is illustrated by employing orthogonal decomposition based on Chebyshev polynomials [122] in two of the examples and on principal component analysis [131] in the third example. In figure 4.2 during the approximate Bayesian computation the measurement uncertainty in the feature vector space is characterised by drawing samples from the posterior distribution in a process of statistical inference.

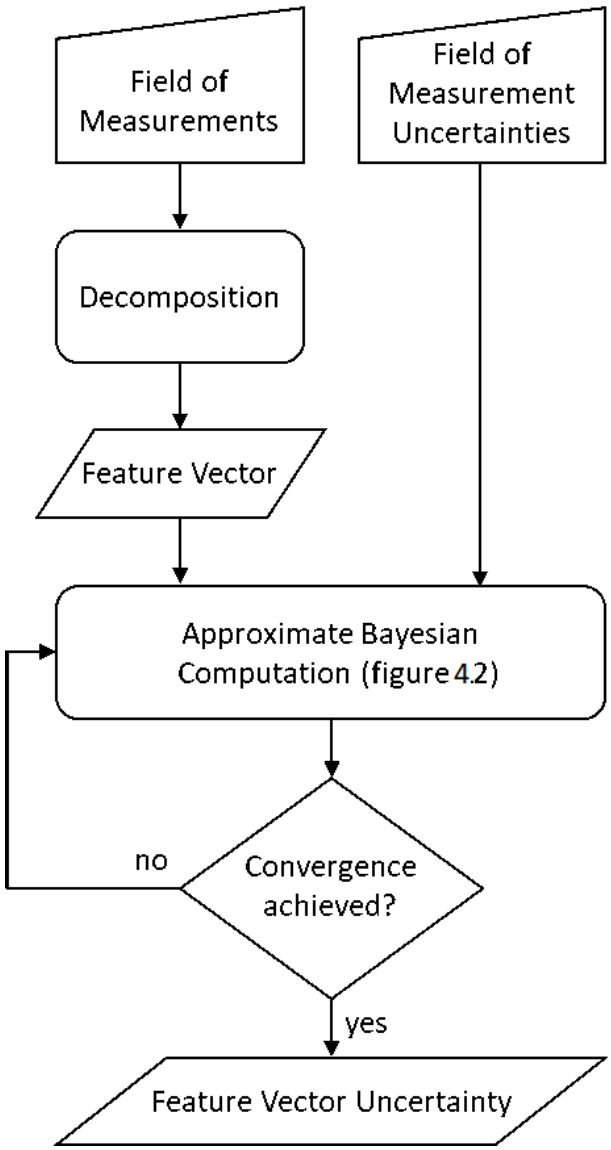

Figure 4.1: Flowchart for estimating the uncertainty in a feature vector representing a field of measurements when the measurement uncertainty is known.

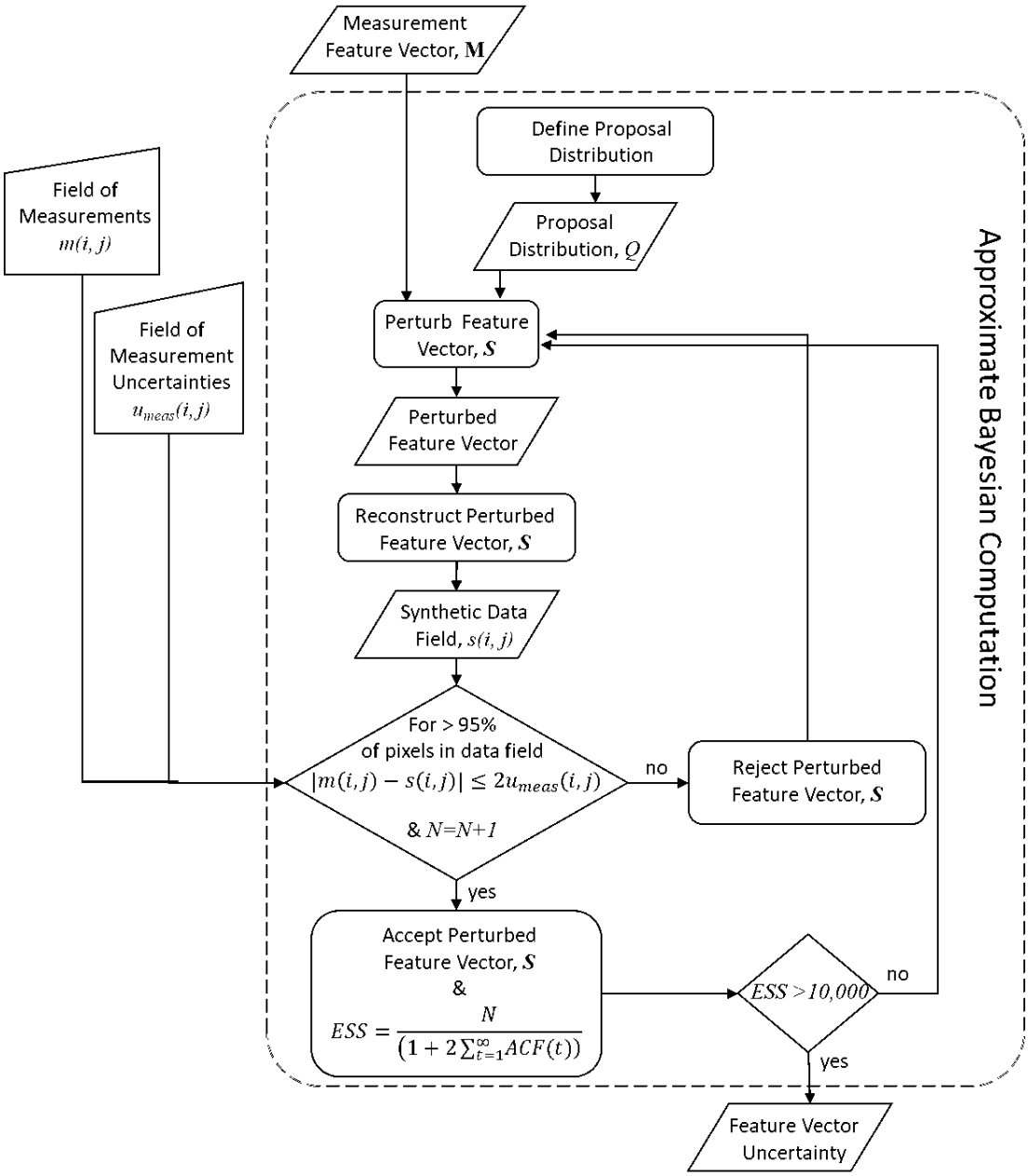

Figure 4.2: Sub flowchart illustrating detail of approximate Bayesian computation ation process shown in figure 4.1.

The approximate Bayesian computation is a relatively new technique developed to allow a posterior distribution to be estimated without knowledge of the likelihood function [31],[132]. The likelihood function is the probability of an event occurring or, in this case, a measurement field being accurately represented by a set of coefficients in a feature vector. In Bayesian analysis, a likelihood function is used to update a prior distribution to generate a posterior distribution,

i.e.

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta} \tag{4.1}$$

where $p(\theta|D)$ is the posterior distribution given the data D, $p(D|\theta)$ is the likelihood function and $p(\theta)$ is the prior distribution, while the denominator $\int p(D|\theta)p(\theta)d\theta$ is a normalizing factor known as the marginal likelihood or evidence. Prior refers to the probability distribution that is assumed to reflect any previous knowledge about the variable or process being modeled, while posterior refers to the probability distribution updated based on some evidence. In the cases considered below, in the absence of any knowledge, the prior is assumed to be a uniform distribution, defined over the range $[\theta_k \pm (2 * max(|M|))]$ for each coefficient of the feature vector, θ_k; where $max(|M|)$ is the maximum coefficient of θ representing the field of measurements. This means that the prior distribution is centred on each coefficient and its width is equivalent to double the magnitude of the maximum coefficient. This selection is motivated by the fact that the largest portion of the variance within a dataset, which is related to the magnitude of the coefficients of the feature vector, is usually described by a few descriptors in a decaying manner, i.e. a few components make up for most of it while the rest quickly decay to zero. This, combined with the fact that the measurement uncertainty is usually smaller than the variance within a dataset, allows the construction of this interval. Thus the width of $2max(|M|)$ was selected as a rule of thumb. The likelihood function is undefined because the physical processes, by which the measurements are generated, are unknown.

To circumvent this, random samples from the posterior distribution are generated, using a Markov chain Monte Carlo (MCMC) technique, and are compared with the experimental measurements using a distance measure to quantify the difference. A sample is accepted if the difference is less than or equal to the expanded uncertainty in the measurements, $1.96u_{meas}$. This process is repeated until sufficient acceptable samples have been generated to define the posterior distribution, i.e. the measurement uncertainty in the feature vector space. The process is summarised in the flowchart in figure 4.2. The implemented version of the approximate Bayesian computation uses the adaptive Metropolis algorithm

[133] as its search tool in the feature space to iteratively search for feature vectors that, when reconstructed into measurement space, yield synthetic data fields for which less than 5% of the pixels deviate from the measurement data field by less the measurement uncertainty in this space, i.e.

$$\text{at least } 95\% \text{ of s(i,j) conform to } |m(i,j) - s(i,j)| \leq 1.96 * u_{meas}(i,j) \quad (4.2)$$

where $m(i,j)$ and $s(i,j)$ are the fields of measured and synthetic values respectively. The uncertainty in the measured values is allowed to vary spatially within the field of measurements by expressing it as $u_{meas}(i,j)$ and, as in the CEN guide [8], the expanded uncertainty, $1.96u$, is used based on the GUM definition [36]. It should be noted that equation (4.2) would need to be modified in the event that the measurement error at each pixel location was very skewed.

The synthetic data fields are generated by perturbing the feature vector, representing the field of measurements. The perturbation is based on a proposal distribution, Q that is a multivariate Gaussian distribution. The choice of the proposal distribution is critical to achieving convergence to the posterior distribution efficiently because proposals must be neither too close to, and hence highly dependent on each other, nor too far apart so that the synthetic values become unrepresentative of the measurement field. In addition, the proposal distribution affects the starting point for the search and the convergence rate. Hence, the standard deviation of the marginals of the proposal distribution are set initially at one fifth of the absolute value of the maximum coefficient in the feature vector representing the measurement field, $max(|M|)$, with a covariance of zero so that the marginals are assumed to be independent. After 1000 iterations, the covariance matrix is updated using data from these iterations which makes the process more efficient [133]. In the event that after these initial iterations, the algorithm fails to find any perturbed feature vectors that are acceptable, then the standard deviation of the marginals in the proposal distribution is reduced until progress towards convergence is observed. There are various measures that can be used to assess the convergence of a Markov chain Monte Carlo method to a stationary posterior distribution. These include, amongst others, the Gelman-

Rubin statistic [134], the Brooks-Gelman-Rubin statistic [135] and the effective sample size (ESS) [136]. The latter was used throughout the examples described in this chapter and is defined as:

$$ESS = \frac{N}{(1 + 2\sum_{t=1}^{\infty} ACF(\tau))} \tag{4.3}$$

where N corresponds to the number of iterations within each Markov chain, while τ represents the time lag used for the calculation of the autocorrelation function, ACF.

The autocorrelation function is an extension of the statistical correlation and is often used in signal processing to assess the similarity of a time-evolving signal with itself across varying time lags τ. The ACF is mathematically described by equation 4.4 where μ is the mean of the random variable, σ^2 is its variance, $E[]$ is the mathematical expectation and t is an integer corresponding to the discrete-time process [137]. As it extends the known statistical correlation it should be stated that it is an indicator of the degree of linearity across the signal and its temporally shifted version across varying time lags. Also its range is limited between -1 and 1, the former suggesting perfect anti-correlation and the latter perfect correlation.

$$\rho_{xx}(\tau) = \frac{E[(X_t - \mu)(X_{t+\tau} - \mu)]}{\sigma^2} \tag{4.4}$$

A rule of thumb suggested by Kruschke [46] for the cases where accurate and stable posterior distributions are sought, is an ESS of 10,000. This makes sure that at least 10,000 of the total iterations, N are independent and representative values of the posterior distribution. The search was conducted three times for each dataset starting from different random starting points in order to ensure that the results are independent of the starting point [46],[138].

Those perturbed feature vectors that satisfy the condition in equation (4.2) represent a cloud of points in feature space which characterise the measurement uncertainty in the space. If the feature space is two-dimensional, i.e. it consists of two components, as in one of the data sets in the first example, where two components suffice to accurately represent the displacement measurements, then

the samples of the posterior distribution can be plotted on a simple graph, as in figure 4.4; however, for a multi-component space, a graphical representation is problematic. Nevertheless, scatterplot combinations can be used as an aid to visualize the drawn samples in multi-component space and can be equally employed to make decisions about the similarity between data sets in model validation, as will be demonstrated in the following chapter, in model updating or in identifying changes in the condition of a system.

At this point the steps needed to implement the Metropolis algorithm for the execution of the approximate Bayesian computations will be explained [132]

- Consider data D corresponding to the measured data field that has been generated from some physical process and is characterized by parameters θ_E. The prior distribution $\pi(\theta_E)$ for each of these parameters that comprise the feature vector is in this case a uniform distribution defined in the range of $[\theta_{E_i} - (2 * max(|M|), \theta_{E_i} + (2 * max(|M|)]$ where the index i corresponds to each of the shape descriptors.

- Generate θ from π

- Reconstruct $D^{'}$ using the generated feature vector θ and calculate the distance $\rho(D, D^{'})$ between D and $D^{'}$

- Accept θ if $\rho \leq \epsilon$, where ϵ is the defined acceptance threshold and return to the first step.

Practically this set of steps is implemented using the Metropolis-Hastings algorithm described below:

1. Generate θ from π

2. Propose a move from θ to $\theta^{'}$ according a proposal distribution $Q(\theta \rightarrow \theta^{'})$ which in this case is a multivariate Gaussian distribution with adaptive step

3. If $\rho(D, D^{'}) \leq \epsilon$ go to step 4. Otherwise stay at θ and return to step 1

4. Calculate $h = h(\theta, \theta^{'}) = min(1, \frac{\pi(\theta^{'})Q(\theta^{'} \rightarrow \theta)}{\pi(\theta)Q(\theta \rightarrow \theta^{'})})$

5. Accept the new set of parameters $\theta^{'}$ with probability h, otherwise stay at θ and return to the first step.

It is also worth mentioning some of the parameters that were used during the execution and post-processing of the approximate Bayesian computation. A commonly used parameter/feature of Markov chain Monte Carlo implementations is the 'burn-in'. This parameter represents the number of iterations early in the Markov chain simulation that are discarded from the final set of parameters. The reason behind that is that as the algorithm is initially settling down to an optimum proposal kernel the initial proposals may be highly correlated and thus unrepresentative of the posterior stationary distribution. The burn-in used in low-dimensional examples (i.e. between 2 and 10 dimensions) was 2000 while for higher dimensionality problems it was about 3000-5000 iterations. Another commonly used practice to ensure low levels of autocorrelation within the Markov chains is 'thinning'. Thinning refers to the process of discarding accepted parameters θ_E every $N-th$ step to reduce the amount of autocorrelation within the chain. Thinning was not used during the runs as it was evidenced that autocorrelation was kept at low levels.

4.3 Applications

4.3.1 Bending displacements in a structural beam

The first example is a simple I-beam with a series of holes in its web (the vertical slender section of the beam) and subject to three-point bending [9]. Two regions of interest have been selected in the web of the beam. In the first, the displacements can be described by a feature vector containing only two components, which renders the explanation of the method and graphical presentation of the results relatively straightforward. In the second region of interest, nine components were required in the feature vector to achieve an acceptable representation of the displacement field. The data were obtained by Lampeas et al. [9] based on a test designed as part of an inter-laboratory study [139]. As shown in figure 4.3, an aluminium I-beam of length 0.5 m and overall cross-section 42x65 mm with flanges (the two horizontal parts of the I) and web (vertical part of the I) of thickness 2.5 mm rested centrally on two supports that were 450 mm apart. The beam was loaded by moving the supports upwards so that contact occurred

between a loading nose situated at the mid-point on the top of the beam. A speckle pattern had been spray-painted onto the web of the beam which allowed the displacements of the surface of the web to be tracked in three-dimensions using stereoscopic images acquired using a pair of CCD cameras that belonged to a commercially available digital image correlation system (Aramis 5M, GOM GmbH, Braunschweig, Germany). In this example, measurements of the displacement in the y-direction, i.e. the direction of the applied load, for two regions of interest shown in figure 4.3, were utilised. Lampeas et al. [9] found the minimum measurement uncertainty to be 0.01 mm using a calibration procedure recommended by the CEN guide [8] and, in this analysis, it has been assumed to be constant throughout the field of measurements.

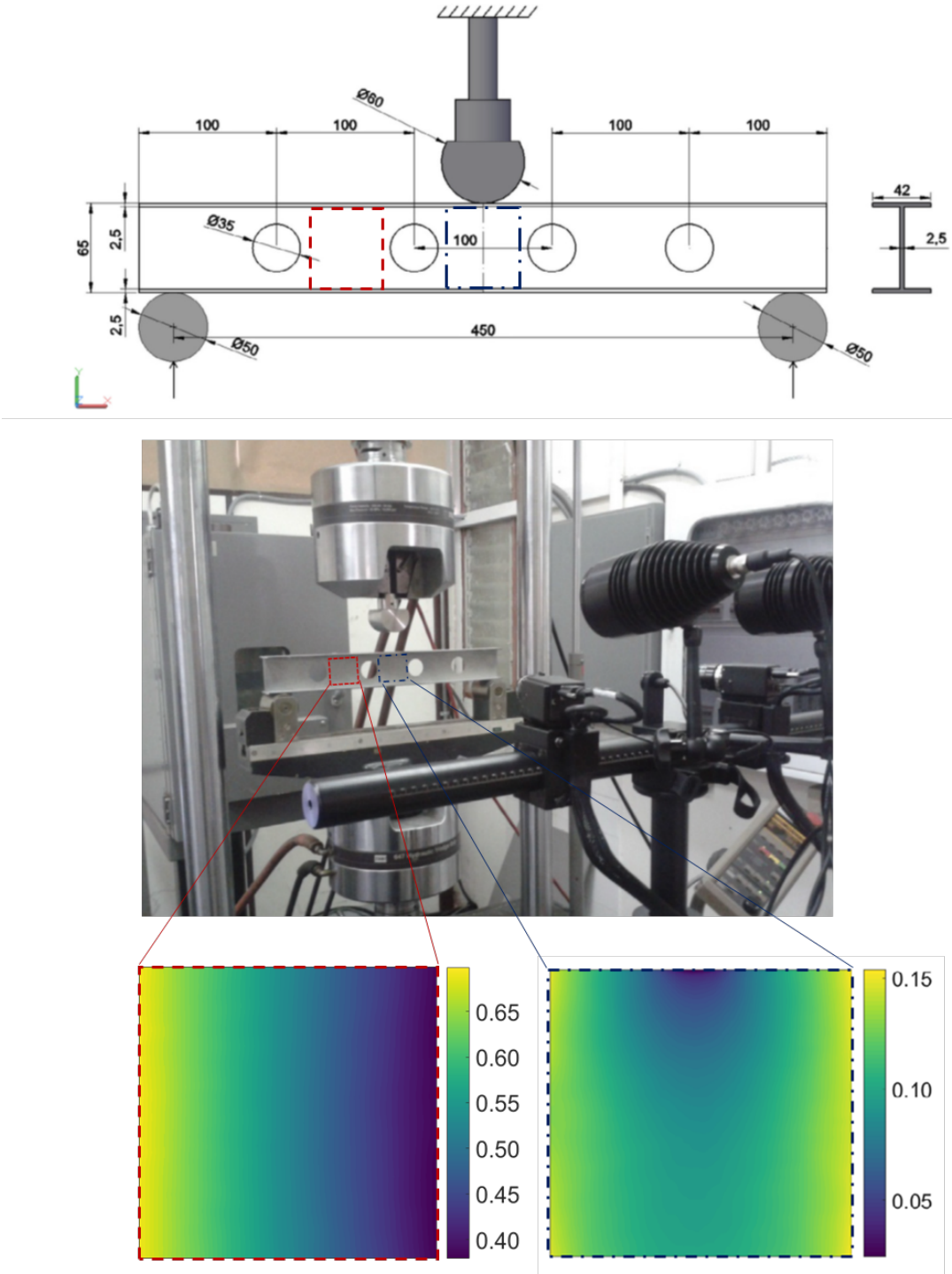

Figure 4.3: Experimental details for first example showing the geometry and loading arrangements for the I-beam (top); the measurement set-up with the digital image correlation system in the foreground (middle); and the vertical (y-direction) displacements of the two regions of interest in the web (adapted from Lampeas et al. [9]).

In the first two examples, i.e. the structural beam and the soil moisture data, the fields of measurements were decomposed into feature vector space by fitting a set of modified orthogonal polynomials and forming a vector from the result-

ant coefficients of the polynomials. A number of suitable types of polynomials are available, including Hahn, Krawtchouk, Legendre and Zernike; each exhibiting a unique set of characteristics amongst which is the sensitivity to local or global features, the type of the domain onto which the data are defined (polar or cartesian) and the form of the polynomials (continuous or discrete) used [140]. Chebyshev polynomials have been used extensively and were adopted here; in part, because the decomposition process could be implemented using a downloadable software that had been prepared for the inter-laboratory study [139] and was readily available [141]. The decomposition process was initially performed using Chebyshev polynomials with a very large number of coefficients. The CEN guide for the validation of computational solid mechanics models [8] recommends that the goodness of fit of the reconstruction of a data field to the original field should be assessed using the average squared residual and that the average residual should not be greater than measurement uncertainty, u_{meas} obtained from a calibration of the measurement system; and further there should be no clusters of residuals greater than three times the average residual, where a cluster is defined as a group of adjacent elements comprising 0.3% or more of the total number of values in the data field. In this example, for the first region of interest, towards the one end of the beam, only the first and third shape descriptors had significant values, i.e. substantially non-zero values, and the reconstruction using only two coefficients gave an average residual that satisfied the conditions recommended in the CEN guide.

The approximate Bayesian computation, described by the flowchart in figure 4.2, was implemented in a specially written algorithm in MATLAB, which was based on one provided by Picchini [142]. The prior was a uniform distribution centered on the feature vector representing the measurement field with a half-width equal to two times the magnitude of the largest coefficient. The results can be seen in figure 4.4, which shows the drawn samples (grey circles) representing the perturbed feature vectors, obtained using the flowchart in figure 4.2, whose reconstructed data fields are different from the measurement field by less than the expanded measurement uncertainty. The feature vector representing the measured data field is shown as a black square and the resulting

grey area represents the uncertainty defined by the measurement error. Figure 4.5 provides evidence of the convergence of the search algorithm to a stationary bivariate posterior distribution with traces of the values of the shape descriptors that the algorithm explored and accepted during the search, the corresponding autocorrelation that should be close to zero, and the frequency distribution for the occurrence of each value of the shape descriptor during the approximate Bayesian computation, which is known as the posterior marginal distribution. The quick decay of the autocorrelation function to zero demonstrates that each step was independent of its predecessor which allowed convergence to be achieved quickly.

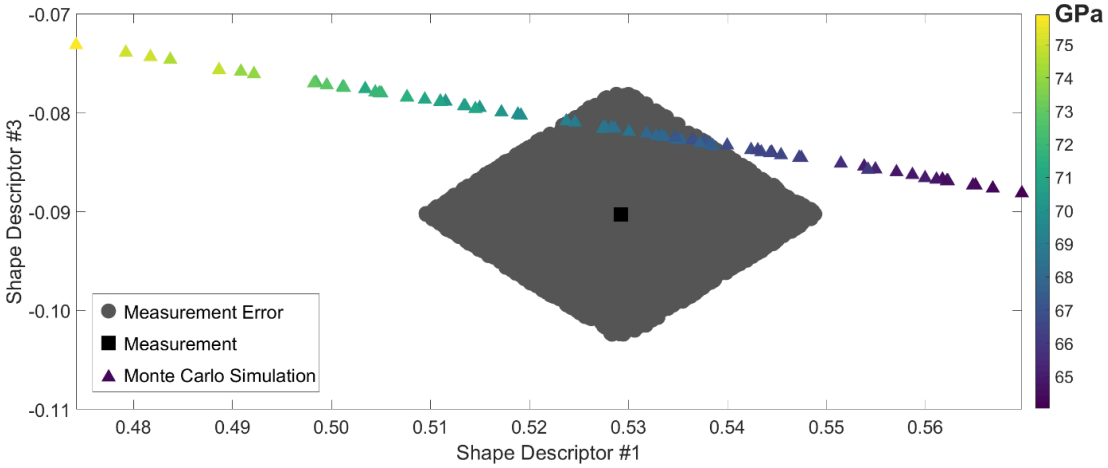

Figure 4.4: Cloud of points (grey circles) representing the perturbed feature vectors, obtained using the flowchart in figure 4.1, whose reconstructed data fields are different from the measurement field by less the expanded measurement uncertainty ; where the measurement field can be described by a feature vector (black square) formed by only two shape descriptors. The corresponding shape descriptors for predictions from Monte Carlo simulations (triangles) based on a range of values of Young's Modulus indicated by the colour bar. The data shown relates to the surface displacement field, in the direction of loading, for the region of interest in the web towards the end of the beam subject to three-point bending shown in figure 4.3.

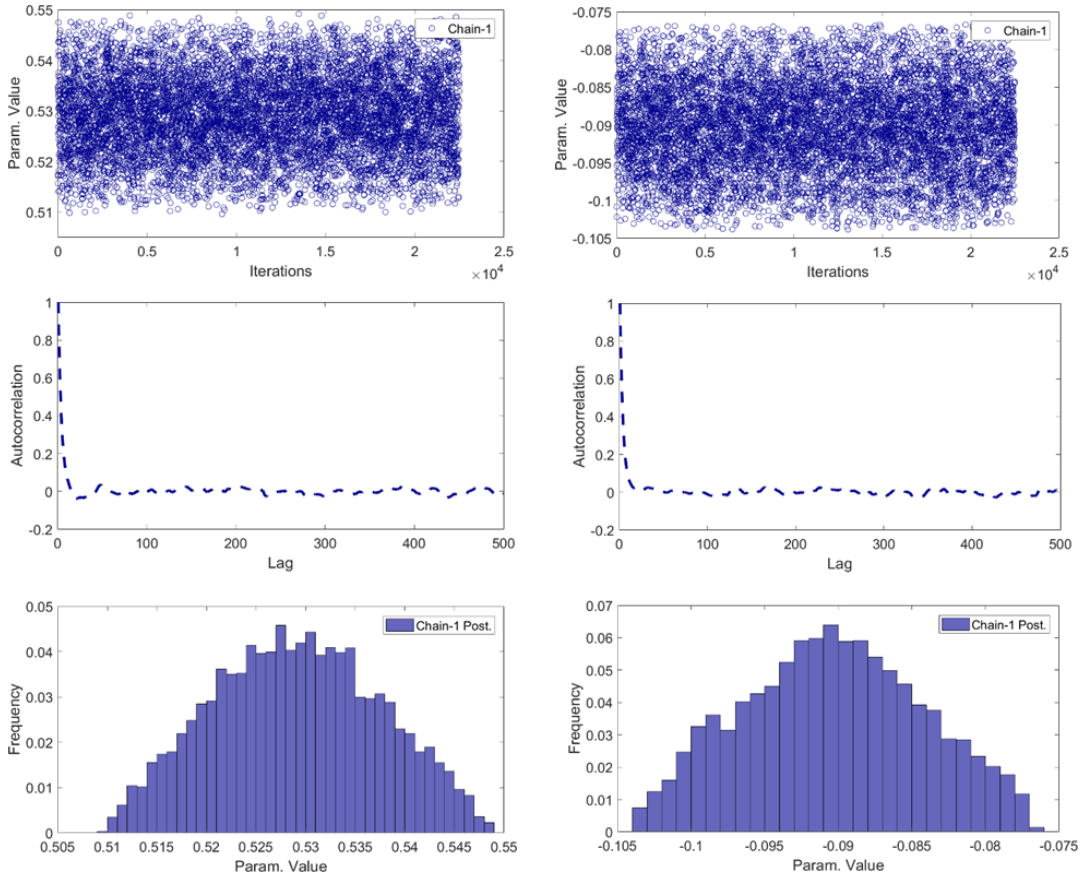

Figure 4.5: Evidence of convergence to the posterior distribution of the algorithm in figure 4.1 for shape descriptors #1 (left) and #3 (right) shown in figure 4.4. The path followed by the search is shown in the top graphs, the autocorrelation (middle) and the posterior distributions (bottom).

The second region of interest was in the web directly under the loading nose. This is a more complicated dataset as shown in figure 4.3, which required decomposition using nine shape descriptors to satisfy the reconstruction criterion specified in the CEN guide. However, the samples of only the three shape descriptors with largest magnitude, namely #1, #6 and #2, drawn during ABC are plotted in figure 4.6. Following the same convention as in figure 4.4, the measurement along with its uncertainty is shown in this three-dimensional plot. The feature vectors that the algorithm visited and found that their reconstructed data fields were to be different from the measurement field by less than the expanded uncertainty, satisfying equation (4.2), are shown.

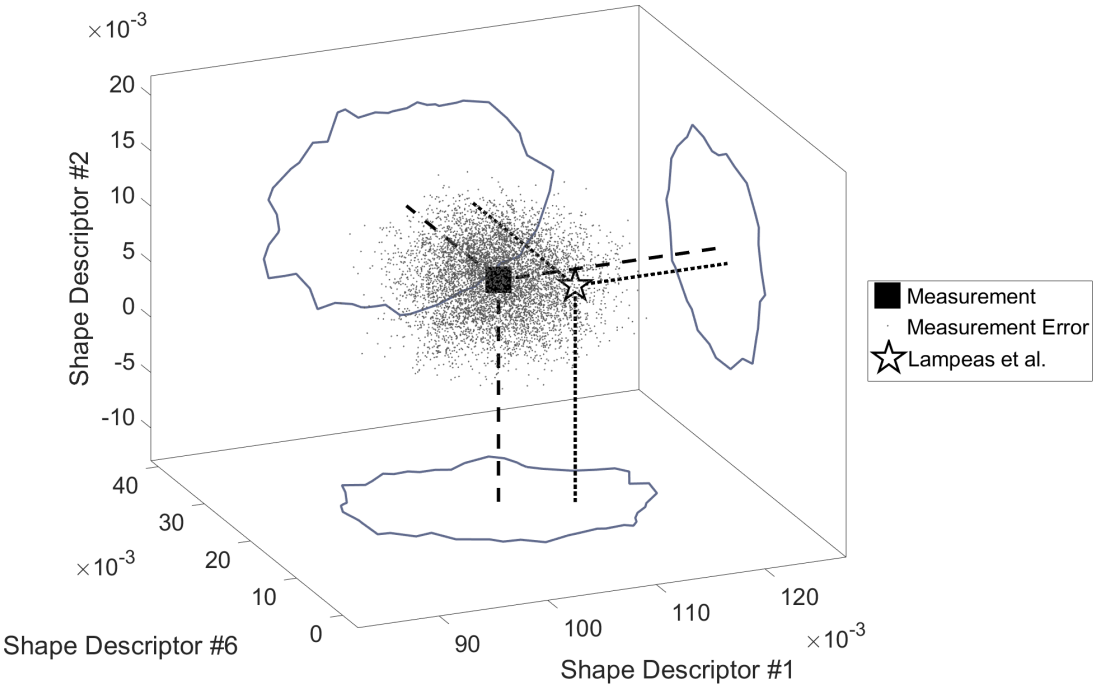

Figure 4.6: Uncertainty bounds for the y-direction displacement in the region of interest directly under the loading nose in the I-beam shown in figure 4.3 based on the three shape descriptors with largest magnitude that represent 99% of the total variability in the measurement data. The cloud of points represents the samples drawn from the posterior distribution, obtained using the flowchart in figure 4.2, corresponding to the measurement and its uncertainty. The prediction by Lampeas et al. [9] lies within the could as demonstrated by its projection and the outline enclosing the points.

4.3.2 Moisture measurements at the Heihe River Basin

Soil moisture data were used for the second example in which the measurement uncertainty varied spatially. The data for soil moisture shown in the top left of figure 4.7 represent the results of a Kriging analysis based on measurements from a wireless network of 162 ecological and hydrological sensors arranged non-uniformly in the Heihe River Basin in China [128]. Three different types of sensors with different measurement errors were used in the study. The variances from the Kriging analysis are shown in the top right of figure 4.7 and account for sparsity of sensors and the heterogeneous measurement error. The proposed method combining orthogonal decomposition and the approximate Bayesian computation was implemented using the dataset on the top left as the measured quantity and the dataset on the top right as the field of uncertainties. Due to the complexity of the measurement field, the initial decomposition was performed using 1000 coeffi-

cients in the Chebyshev polynomials and then the 100 largest non-zero coefficients were retained as elements in the feature vector in order to satisfy the requirements for quality of the representation recommended in the CEN guide [8].

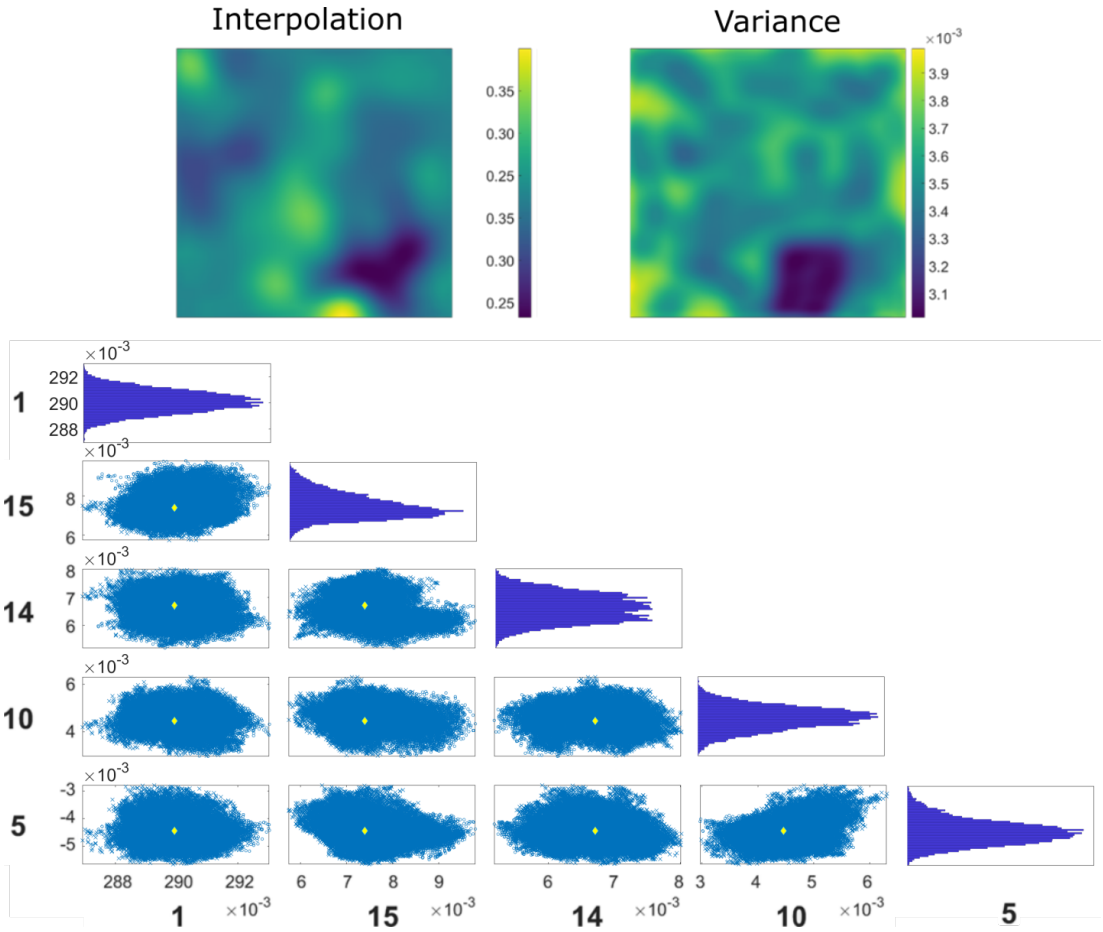

Figure 4.7: Spatial distribution of soil moisture data from the Heihe River Basin, digitised using data from Kang et al. [128] based on the results of Kriging interpolation (top left) from sparse measurement locations with heterogeneous measurement errors represented by the Kriging variance (top right); and the corresponding uncertainty bounds (bottom), based on the five most significant shape descriptors. The measurement field is described by a feature vector represented by the yellow diamond. Histograms reflecting the distribution of each shape descriptor are shown in the diagonal.

An unavoidable consequence of the complicated shape of the data field is the large number of shape descriptors required in the feature vector to represent it to the required accuracy. However, it has been shown that the adaptive Metropolis algorithm, used in the approximate Bayesian computation, can efficiently handle searches in such high dimensions [133]. Some of the results of the search are shown at the bottom of figure 4.7 for combinations of the five of most significant shape descriptors. The array of plots represents an attempt to present the five-

component cloud of points that characterise the uncertainty for the measurement field in the low-dimension or feature vector space. This multi-variate 'cloud' of points corresponding to the samples drawn from the posterior distribution cloud could be used to assess, for example, the significance of changes in the soil moisture over a time period.

4.3.3 Monthly oceanographic temperature fields

In 2000 a global network, currently consisting of about 3800 Argo profiling floats, was established with the aim of systematically observing the temperature and salinity of the world's oceans. The resultant high quality and spatially dense data has allowed researchers to obtain a better image of the properties of the world's oceans and their interaction with climate changes. The information gained from the network is being used to drive policy changes related to climate and also to validate climate models [143].

However, before the data can be used for climate research, quality flags are attributed to the measured quantities and those that pass the quality requirements are assembled through a process of optimal interpolation into plotted fields, such as the one shown in figure 4.8 that is based on the In Situ Analysis System (ISAS) 13 for which more details can be found in reports by Gaillard and his co-workers [129],[130]. The results of this interpolation process are monthly averaged temperature and salinity fields across the globe along with fields of errors that are based on four components: (i) the measurement error of the floats; (ii) the variance in these fields measured within a time frame of 41 days with respect to the mean; (iii) the uncertainty arising from the interpolation process; and (iv) previous statistical knowledge in parts of the ocean where measurements are scarce and estimates are provided by previous analyses.

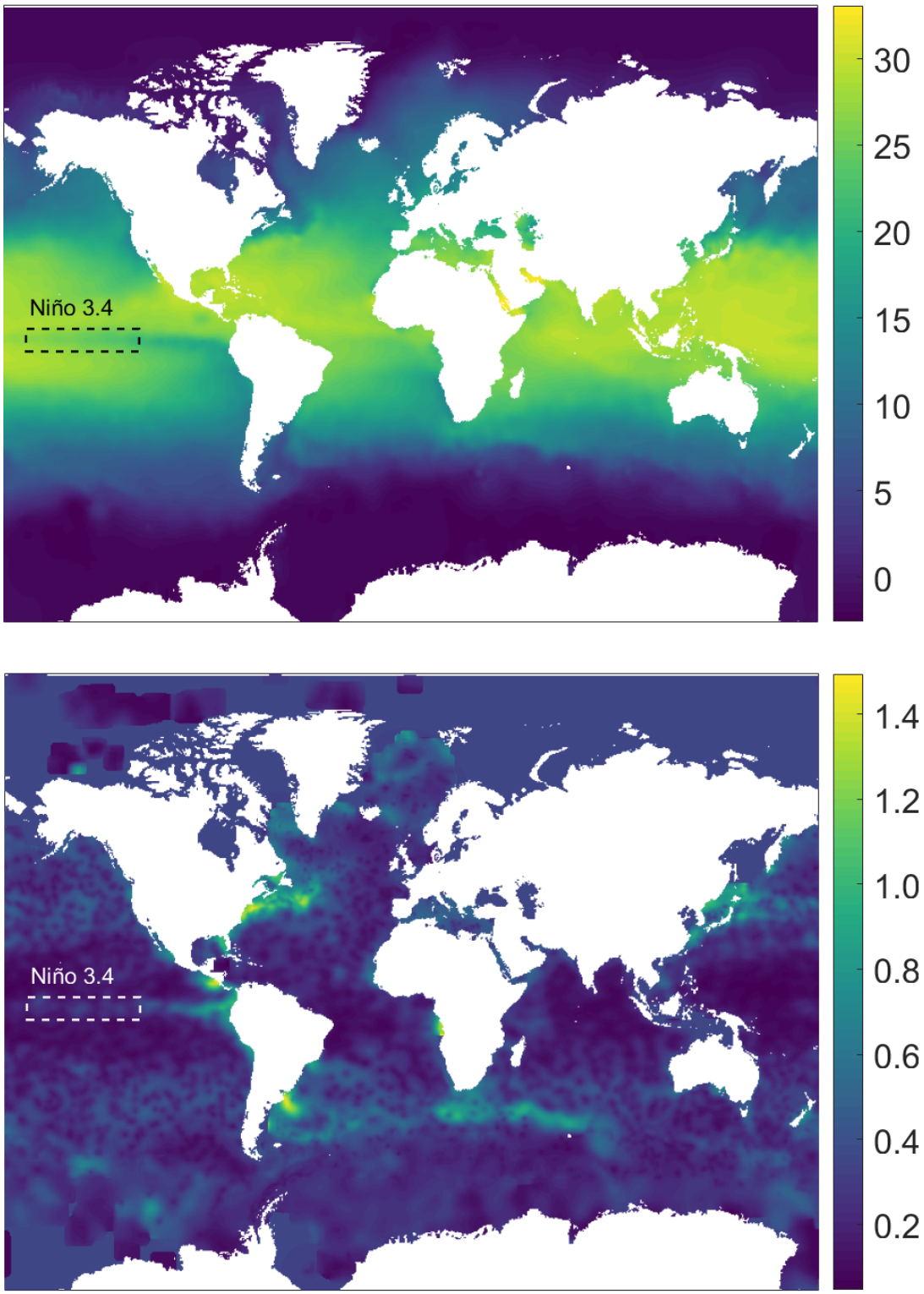

Figure 4.8: Monthly ocean temperature (°C) distribution (top) at a depth of 10m for September 2007 and corresponding error field (°C) (bottom) from Gaillard [130]. Niño region 3.4 is shown in the dashed rectangles.

The data fields used in this example are monthly temperature data spanning a total of 11 years from 2002 to 2012 from Gaillard [130] and the illustrative data

in figure 4.8 is for September 2007. PCA was used to decompose the monthly temperature data. PCA allows the projection of high dimensional data into a lower dimensional space by retaining only the coefficients of the components that account for the largest percentage of the total variability in the data [131], [144]. This results in a set of uncorrelated orthogonal basis vectors, each representing a certain feature or mode of the dataset and a set of coefficients. The dataset can then be reconstructed as a linear combination of the basis vectors and the corresponding coefficients, i.e. the outcome is similar to decomposition using orthogonal polynomials though the process is different with the result that features or modes are dependent on the form of the original dataset whereas they are fixed in the polynomial decomposition. The analysis involved a number of steps: initially the 132 monthly temperature fields, consisting of the same sized matrices for each month, were reshaped into vectors, after the gaps in the datasets representing land masses were removed leaving 270,733 values in each vector. Second, these vectors were assembled into a large 270,733 x 132 matrix in which each vector formed a column. Thirdly, the matrix was centred around its mean and decomposed using principal component analysis to generate a matrix of coefficients of the principal components, with a feature vector for each month, and a matrix of eigenvectors corresponding to the principal components, also known as the basis vectors. The complexity of the ocean temperature distributions required 100 principal components to describe them so that the root mean square error (RMSE) of each reconstructed dataset compared to its corresponding original dataset was always less than the mean uncertainty in the temperature measurements.

Finally, approximate Bayesian computation (figure 4.2) was performed using as inputs the feature vectors and the monthly fields of uncertainties, while the principal components were used for the reconstruction of the perturbed feature vectors. The prior was a uniform distribution centered on the feature vector representing the measurement with a half-width equal to two times the absolute value of the first principal component, in order to minimize the effect of the prior on the posterior distribution. The results are shown in figure 4.9 for September 2007 using temperature and error fields from an ocean depth of 10m for the ten most significant principal components. The cloud of points represents the

uncertainty bounds on the temperature data in the feature vector space and can be used to evaluate the significance of trends in the data in this space.

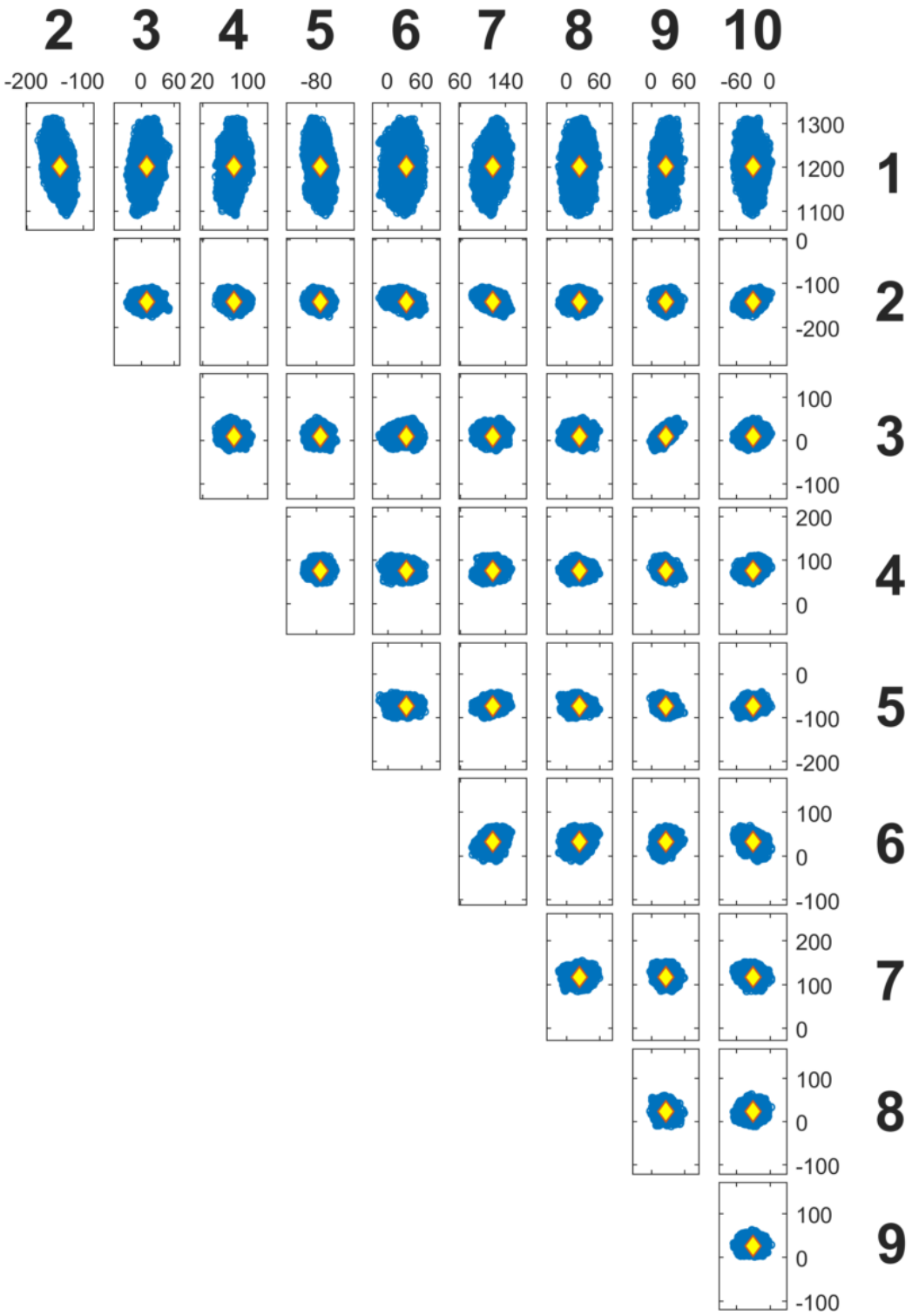

Figure 4.9: The distribution of measurement error in the feature vector space for the first ten principal components for the oceanographic data. The measurement is depicted by the diamond at the centre of each plot.

4.4 Discussion

4.4.1 Representing the measurement error

The objective of this work is the development of a method to characterize the uncertainty associated with spatial measurements in a low-dimensional form. The significance of this method is that it can be used in cases where information about the measurement error form is limited. Various types of mathematical transformations can be used to extract features or patterns from the data to reduce its dimensionality [145], [140]. In various applications, where the associated uncertainties are not negligible and the decision-making process is based on a representation of the data in a lower dimensional form, it is important to be able to assess whether a pair of feature vectors belong to the same population. To be able to carry out this assessment, it is required that the associated uncertainty be accurately represented in the feature vector space. In this chapter, it has been proposed that the extent of the uncertainty in feature vector space can be established using approximate Bayesian computation via an adaptive Metropolis algorithm to search for perturbed feature vectors that, when reconstructed into measurement space, generate synthetic data fields that deviate by less than the expanded measurement uncertainty from the measured data field. The set of such perturbed feature vectors represent the uncertainty in feature vector space. In a two-dimensional space, such as in the first region of interest in the first example, this set of perturbed feature vectors can be readily represented in a graph such as in figure 4.4; however, when the feature vector space involves many components then sets of scatter-plots, such as those in figure 4.9, can be used to represent the multi-component uncertainty.

Propagating the measurement uncertainty in temporally evolving measurements when the form of the error is known (traditionally assumed normal) and uncorrelated with the underlying signal, usually involves sampling from the error distribution and then adding it to the signal within the context of a crude Monte Carlo simulation. However, when the error structure is more complex, as can be the case in spatial measurements, where the presence of spatial autocorrelation requires numerous assumptions to accurately model the underlying process, the

efficacy of such crude simulations is limited. An example of that limitation, for the case of uncorrelated Gaussian noise, is described with the aid of figures 4.10 and 4.11. Figure 4.10 (left side) depicts the linearly varying u_y dataset where 2 shape descriptors were used to accurately represent it in the feature vector space. Afterwards, uncorrelated Gaussian noise was added $X \sim \mathcal{N}(0, u_{meas})$ to the raw measurement (right side of the same figure) and the resulting dataset was decomposed. This process was repeated 10,000 times and the result is shown in figure 4.11.

The conclusions that can be drawn from this case are the following: the effect of high frequency, uncorrelated noise, is quite limited, indicated by the concentration of the decomposed datasets in the inset of figure 4.11. This is not a surprise as the impact of such high-frequency spatial characteristics is minimal on the two shape descriptors selected to describe the measurement in the feature vector space. Even if more shape descriptors were used to capture the variability attributed to that form of uncertainty, it should be reminded that the results would be founded on the assumption of Gaussian, uncorrelated error. However, this is not the objective of the proposed technique where the aim is to minimize the number of assumptions related to representing that error whose form is usually unknown.

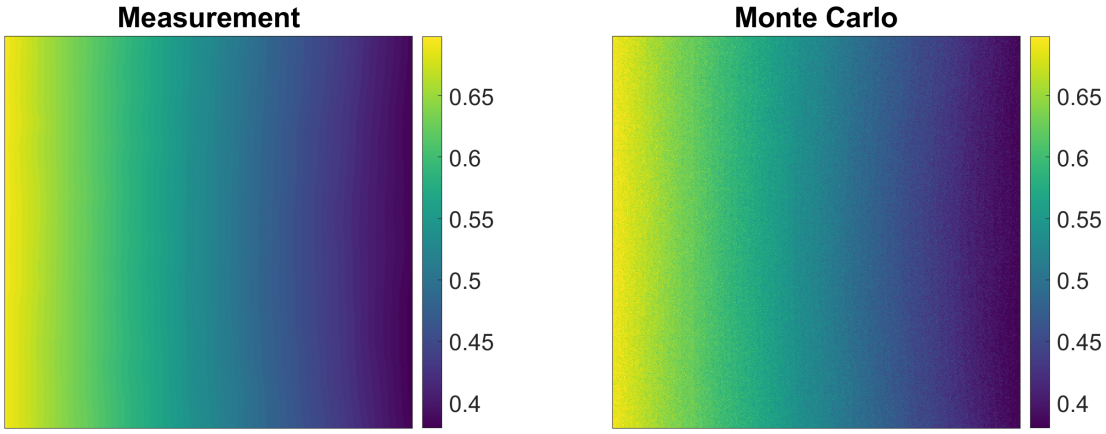

Figure 4.10: The u_y displacement measurement corresponding to the left side of the I-beam, as depicted in figure 4.3, is shown on the left side, along with uncorrelated measurement error on the right side.

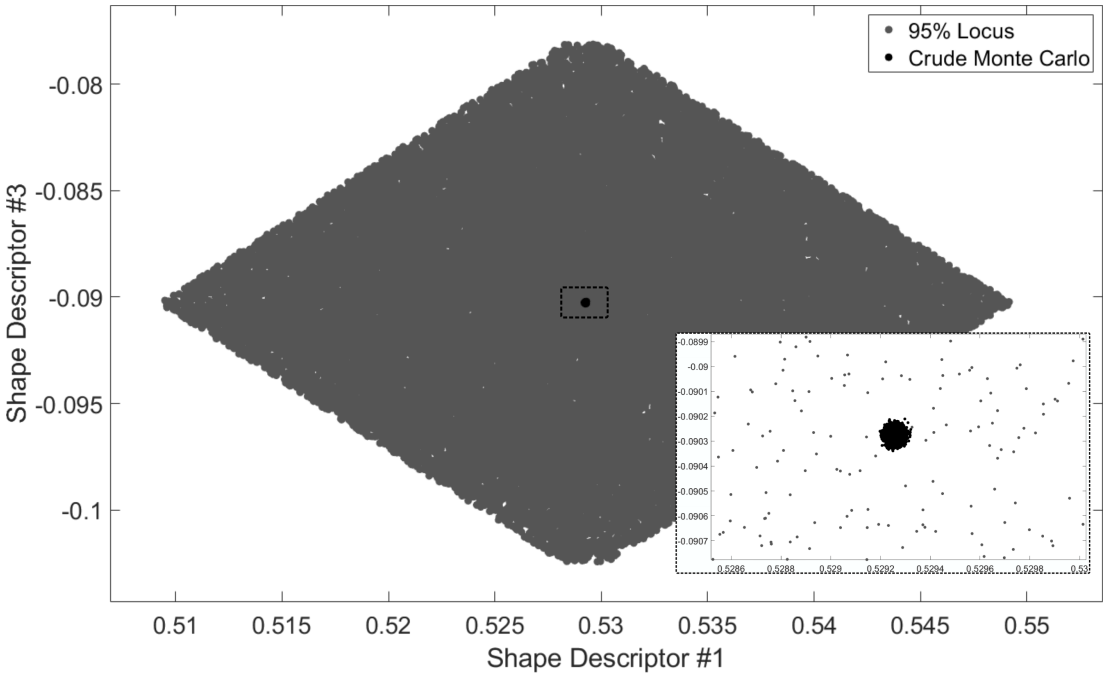

Figure 4.11: Reproduction of figure 4.4 along with the results of decomposed Monte Carlo simulations corresponding to uncorrelated measurement error in the inset

It should be highlighted that Monte Carlo techniques are not limited to cases where the error is uncorrelated and Gaussian. To explain this statement the focus is shifted to geostatistics where different research workers have focused on the problem of accounting for the presence of error in the measurements. A significant portion of the research work on the field is based on Kriging [81]. This modelling technique allows users to create continuous spatial maps of an observed quantity while utilizing information from multiple measurements. This is achieved as a combination of three main components: a model that approximates the covariance structure across the measurements, knowledge about the distribution of the underlying process and knowledge about the form of the measurement error. Afterwards, Monte Carlo simulations can be utilized to draw realizations of the underlying spatial process.

For example, Kang et al. [128] developed a non-parametric bootstrap method to account for the presence of heterogeneous measurement error in measuring devices sparsely located across the Heihe River Basin. Their technique, building on the one developed by Christensen [146], similarly to most of the research in geostatistics is based on the assumption of stationarity [81], which may not always be well-founded. On the other hand, non-parametric spatial bootstrap

techniques developed by García-Soidán et al. [147] and Castillo-Páez et al. [148] allow for more accurate calculation of estimators from the underlying data, while limiting the number of assumptions. Even though their work tackles many of the problems associated with obtaining accurate estimators from spatially correlated data, it does not handle the presence of measurement errors.

Another issue associated with modelling the spatial correlation using geostatistics-based methods is the high computational cost involved during the construction of the covariance matrix. This can be prohibitive when thousands of grid point measurements are available. Although geostatistics-based techniques could provide viable alternatives in the process of representing the measurement error in a lower-dimensional space, the underlying assumptions needed to hold true make them less attractive. It is apparent that Monte Carlo simulations have a wide range of applications and could be used to represent the measurement error in the feature vector space. To do so however, they require information about the form of the measurement error which may not always be available. The proposed method is an alternative to such cases and can be used to represent that error in the feature vector domain while making limited assumptions about its form.

4.4.2 Applications

One of the potential applications of the proposed technique is in model validation. Decision makers want to know whether they can trust the predictions made by models across science and engineering, from mechanics and meteorology to climate modelling and finance. Part of the process of establishing trustworthiness is to perform a validation process, which has been defined as 'determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model' [4]. This is not straightforward when fields of measurements and predictions are available, particularly when the data fields have different grid densities, orientation and scales. Thus, to alleviate these issues in structural mechanics, the CEN guide for validation of computational models in solid mechanics [8] recommends reducing the dimensionality of the data fields using orthogonal decomposition by employing suitable polynomials, as described for the first example. However, once the data are reduced to

a lower-dimensional space, a rigorous representation of the associated measurement uncertainty is seldom made. Instead the CEN guide recommends plotting the shape descriptors, describing the measured and predicted data fields, against one another and assessing whether the resultant points lie within an interval defined by

$$\{\mathbf{s}_M\} = \{\mathbf{s}_E\} \pm 1.96u_E \tag{4.5}$$

where $\{\mathbf{s}_M\}$ and $\{\mathbf{s}_E\}$ are the shape descriptors representing the model predictions and measurements respectively and $1.96u_E$ is the expanded uncertainty in the shape descriptor describing the measurements which is given by

$$u_E = \sqrt{u_{meas}^2 + u_{RES}^2} \tag{4.6}$$

u_{meas} is the measurement uncertainty obtained from a calibration of the measurement instrument, while u_{RES} is the average residual obtained from comparing the reconstructed and original data fields as mentioned earlier. In the process described in the CEN guide, it is assumed that the measurement uncertainty is uniform over the field of measurements and is not transformed into the shape descriptor space (it is a deterministic comparison between two feature vectors with an accept/reject outcome). However, the proposed methodology allows a graphical representation of the measurement uncertainty in the low dimensional space as a cloud of points with the shape descriptor representing the measurement values at its centre, as shown in figure 4.4 for the region of interest towards the end of the I-beam. In addition, to performing an experiment with the I-beam, as shown in figure 4.3, Lampeas et al. [9] also predicted the behaviour of the beam using a finite element model. The predicted field of displacement for the region of interest in the centre of the beam was decomposed using Chebyshev polynomials in exactly the same way as the measured field and the resultant shape descriptors are plotted in figure 4.6 and lie just within the cloud of points representing the uncertainty interval for the measurements. Thus, it could be concluded that the model is an acceptable representation of the experiment because the difference between the predictions and measurements is less than the expanded uncertainty in the feature vector space, following the same principles as the CEN guide but

applying them completely in the feature vector space. This conclusion agrees with the one that was drawn by Lampeas et al. [9] using the criterion described by equations (4.5) and (4.6).

Calibration

It is also possible to use the uncertainty described by posterior distribution to calibrate a model. For instance, when the finite element modelling for the I-beam is repeated using a series of values for the Young's modulus varying between 65 and 75 GPa then the series of coloured triangles in figure 4.4 represent the predicted displacement fields. Because the values for the Young's modulus that yield shape descriptors within the distribution lie between 67 and 71 GPa it can be concluded that this range would be acceptable when considering the displacements in this region of interest.

Although it is relatively straightforward to develop a computational model of the structural behaviour of the beam in the first example, it is considerably more complicated to construct computational models for soil moisture or for ocean temperatures due to the large number of parameters involved and the complexity of the interactions between factors influencing the responses. In such circumstances, it is often impractical to perform multiple runs of a model thus an alternative is to employ techniques such as meta-modelling to overcome this issue. Meta-models are simplified surrogates for models of the system of interest in which the relationship between the inputs and outputs of the original model are represented mathematically using a technique such as an artificial neural network, polynomial chaos expansion or Gaussian process regression. These techniques can successfully describe the complex mapping between outputs and inputs; however, they do not provide a representation of the associated uncertainty in the corresponding space. The proposed methodology could be used alongside such techniques to accurately represent the uncertainties in the reduced-order or feature vector space. This is effectively the process represented by the example using the soil moisture data from the Heihe River Basin [128] that are based on the results of Kriging interpolation.

Identification of critical changes

It would be expected that the volume of the cloud of points characterising the measurement uncertainty in feature space would be correlated to other measures of the errors in the measurements. This has been shown in figure 4.12 for the ocean temperature data by plotting the volume of the cloud of points representing the posterior distribution as a function of the monthly average errors, i.e. the spatial average of the errors in the dataset for each month. The volume was calculated as the square root of the determinant of the covariance matrix of the distribution. The calculation of the determinant of the covariance matrix as an estimate of the scatter of a multivariate distribution has been reported in various sources such as [149] and [150]. A covariance matrix consisting of ten of the most significant principal components was used in this case and gave a correlation of 0.975 with the monthly average errors; the noise in the volume data is likely a result of the complexity associated with the dimensionality of the problem. The need to characterize the temporally varying uncertainties in measurements could be important as new equipment may be added to enhance the overall credibility of the measurements or to remove damaged sensors.

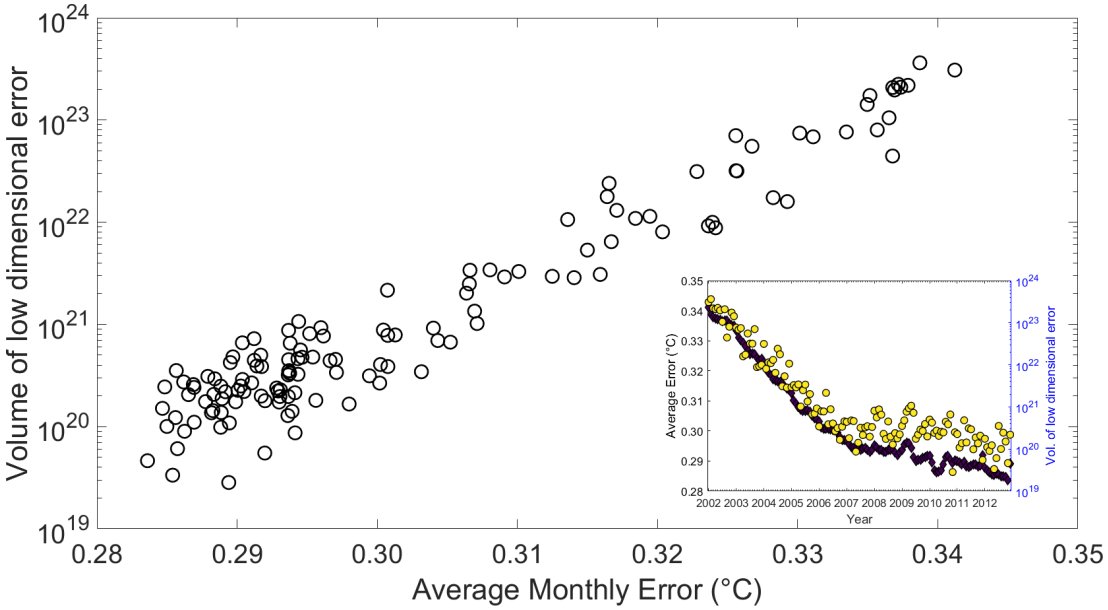

Figure 4.12: The volume of the cloud of points representing the posterior distribution as a function of the monthly average errors, i.e. the spatial average of the errors in the dataset for each month, for the month ocean temperature data from Gaillard [130]; and inset average error (diamonds) and the volume plotted as a function of time.

The volumes defined by the 'cloud of samples' for the temperature data for each month in 2002 are shown in figure 4.13. It can be seen that they are distributed along an approximately elliptical path running clockwise through the year. There is no overlap between the hulls which implies it would be reasonable to conclude that there is a significant difference between the global temperature pattern in each month. A more sophisticated analysis is possible by examining the behaviour of certain shape descriptors or principal components. For instance, it was observed that the fifth principal component (PC-5) describing the monthly distribution of temperature could be used to characterise the El-Niño Southern Oscillation (ENSO) as shown in figure 4.14.

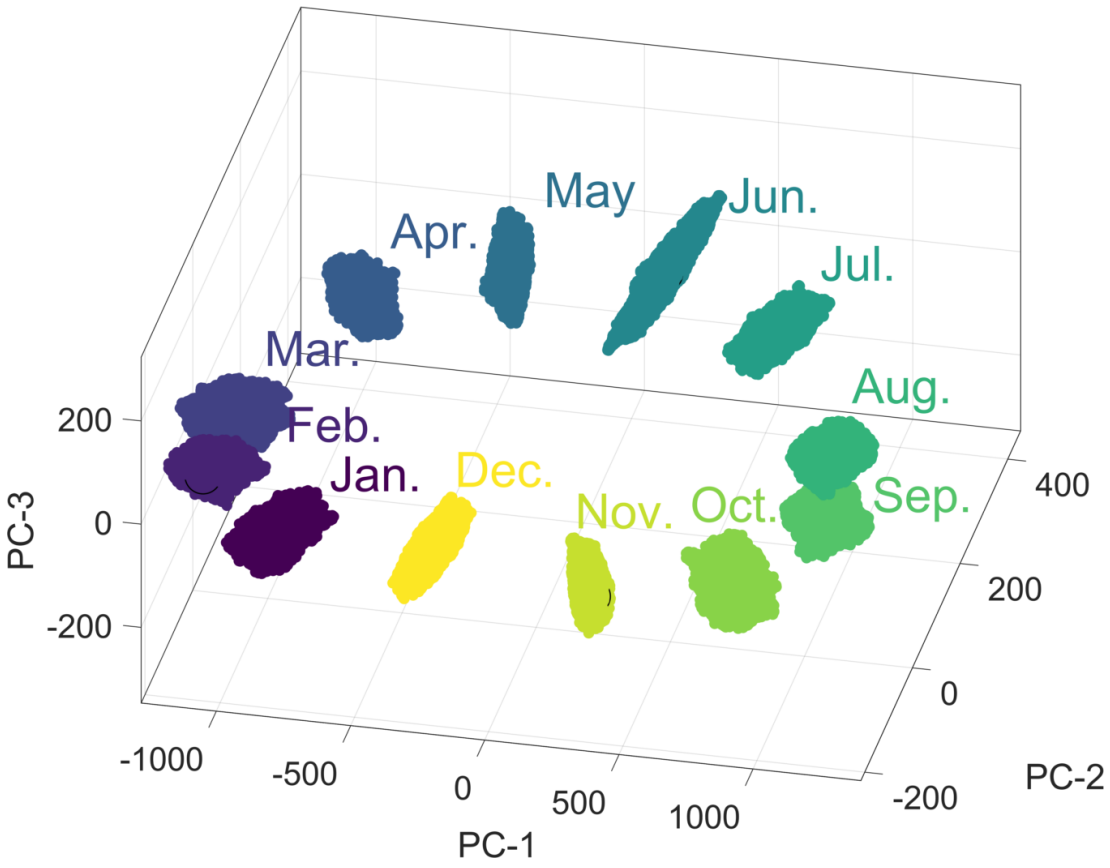

Figure 4.13: The sets of points representing the uncertainty intervals for the ocean temperature measurements for each month in 2002 using only the three most significant principal components. The lack of overlap between sets can be interpreted to imply a significant difference in the temperature between months.

The ENSO is an irregular cycle of recurring warm (El Niño) and cool (La Niña) patterns of temperature in the tropical Pacific that occur every two to seven years and cause major disruptions in the climate [151]. The Oceanic Niño Index (ONI)

is the difference between the three-month average and the 30-year average of the surface temperature of the ocean in an area of the east-central tropical Pacific between 5°N and 5°S and between 120° and 170°W, which is known as the Niño 3.4 region and is shown in figure 4.8 [152]. The correlation between the value of the fifth principal component (PC-5) and the Oceanic Niño Index (ONI) was 0.88, which implies that PC-5 captures the characteristics of the ENSO phenomenon.

The methodology proposed in this study can be used to define an uncertainty interval for each principal component as the distance across the cloud of points in the direction corresponding to each component. It was found that this uncertainty interval can be used as an indicator of an ENSO phenomenon when the value of PC-5 varies from zero by more than its expanded uncertainty for three consecutive months. The variation of the value of PC-5 and its uncertainty interval are plotted in figure 4.14 as a function of time, together with the value of the Oceanic Niño Index (ONI) [153] and the shape of PC-5 is shown as an inset.

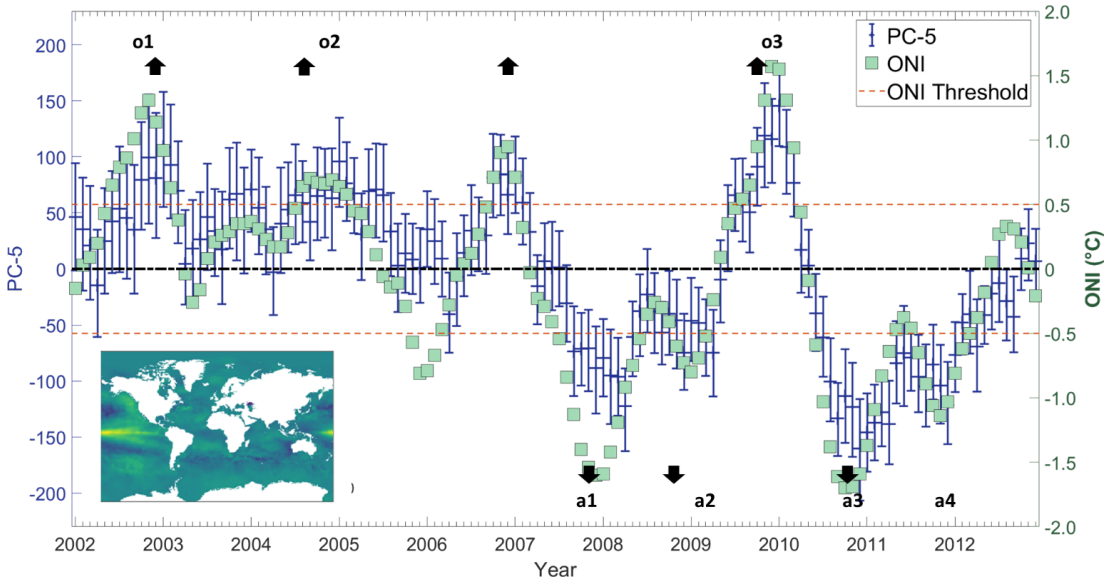

Figure 4.14: The magnitude of the fifth principal component, PC-5 of global ocean temperature at a depth of 10m and the Oceanic Niño Index (from [151]) as a function of time. El Niño and La Niña phenomena based on the ONI are highlighted by o1 to o3 and a1 to a4 respectively; while the corresponding phenomena indicated by the PC-5 varying from zero by more than its uncertainty are indicated by upward and downward errors respectively. The inset shows the shape of the fifth principal component.

The Climate Prediction Center of the National Oceanic and Atmospheric Administration (NOAA) consider that La Niña conditions exist when the ONI is

less than or equal to -0.5 and El Niño conditions when it is greater than or equal to 0.5 for at least five consecutive months. La Niña and El Niño events identified using the ONI criteria are numbered in figure 4.14 using the prefix A and O respectively; while those identified using the PC-5 criteria are shown by upward and downward arrows respectively. It can be seen that the PC-5 criterion predicts all of the ONI indicated events but also gives one false positive in the fall of 2006 when the ONI is over the 0.5 threshold for only four months. In 2011-12, the PC-5 criterion indicates a 20-month La Niña event whereas the ONI implies two events separated by three months. This proposed approach to classifying the occurrence of ENSO phenomena has a number of advantages over the ONI, namely: that the uncertainty interval could be used to provide a level of confidence in the classification; it should be more representative of the mechanisms driving the ENSO phenomena because it is based on the global pattern of ocean temperatures; and it should allow straightforward comparison of predictions and measurements while accounting for the global ocean dynamics. Even though the application based on the El Niño Southern Oscillation does not include a model prediction, its significance in the identification of changes in temporally evolving phenomena is apparent. The use of measurements and predictions for quantitative model validation will be demonstrated in more detail in the next chapter.

4.4.3 Implementation

The reasons for selecting approximate Bayesian computation to search for the posterior distribution of the measurements in the coefficient or feature vector space were: i) its tractability, especially when moving to high-component spaces compared to other techniques such as history matching [82], which would require a much larger number of iteration ii) the rich literature around Markov chain Monte Carlo techniques iii) the potential capabilities for faster convergence using techniques such as adaptive stepping [133] and, iv) the ability to run multiple 'chains' of calculations independently, thus exploiting parallel programming capabilities in modern computing.

The principal benefit stemming from the application of the new methodology is the way in which it supplements existing techniques for reducing the dimen-

sionality of information-rich data fields by allowing the associated uncertainty to be characterised and represented in the low-dimensional space. As the examples have demonstrated, the results from the methodology provide a visual and intuitive way to inform the decision makers about the variability in the data and the significance of the difference between data fields based on rigorous statistical principles. Compared to traditional geostatistical approaches to characterizing the uncertainty in spatial data, where a large number of assumptions and choices must be made during the analysis, the proposed methodology requires only two parameters to be selected, namely the size of the uncertainty interval and the confidence level required (e.g. $1.96u(i,j)$ for a 95% confidence interval) ; hence, the methodology can be used as a 'black-box' approach. This is attributed to the fact that the primary spatial characteristics of the data are captured during the decomposition process and subsequently used to represent the associated uncertainty. The resulting distribution in low-dimensional space, representing the spatial data fields is easier to handle using multivariate statistics thus allowing inferences to be drawn. Finally, employing the proposed methodology allows all of the available spatial information to be included in the analysis. This is important in activities like model validation and model calibration or updating, where all the existing information, including both measurement values and the accompanying uncertainty, should be taken into consideration when making decisions.

Another widely known technique associated with dimensionality reduction is Factor analysis. Factor analysis can be used to describe relations among correlated observed variables as a linear combination of unobserved (latent) variables, called factors, that cannot be directly measured. For example, a person's intelligence cannot be directly measured; it can be inferred indirectly from their responses from a series of questions or puzzles. Similarly, the exchange rate between currencies could be attributed to several factors which cannot be directly measured, including interest rate decisions by central banks, geopolitical influences or prevailing market sentiments about the economic future of the two countries.

In engineering, factor analysis has been used in [154] to remove the impact of environmental variations from damage sensitive features, while in [155] as the

means to identify the onset of damage in a dynamically loaded structure. Even though factor analysis provides the basis for modelling the response of a structure as a function of unobserved common factors it is a pre-requisite that the number of these factors is readily known. This property can also impact the representation of the data in the low-dimensional space as the coefficients (known as the factor loadings) are affected by the number of factors (compared to PCA which are invariant to the number of the selected components) and can lead to different results. In general, latent variable models, of which factor analysis is a member of, are used to model the relation between a series of observed variables as a function of unobserved, latent variables. The significance of their use has been demonstrated in a number of research publications and a method to quantify the uncertainty in the loadings as a result of measurement error could be the basis for future research; for the time being this is out of scope as the aim in this work is to use robust feature extraction techniques to transform the data into a lower-dimensionality space and then inform decisions associated with the quality of predictions.

4.5 Conclusions

A novel methodology has been developed that allows the characterization of the uncertainty of the coefficients of a feature vector representing a field of measurement values with known measurement uncertainties. The method uses approximate Bayesian computation with an adaptive Metropolis algorithm to search for the posterior distribution of the measurement values and their uncertainty in the feature vector space. The result is a distribution in the feature vector space that characterise the measurement and its uncertainty and forms a multi-component uncertainty estimate that can be used to evaluate the significance of differences between data fields.

The innovations in this methodology lie in:

- Its capability to characterize the uncertainty in the elements of a feature vector when the uncertainty in the underlying measurements is spatially constant or varying. The uncertainty may be obtained either from the

calibration process for a single device capable of measurements across a field of view or through statistical post-processing as in the case of spatially dispersed sensors whose error is heterogeneous;

- Its applications to the validation or confirmation of models in engineering mechanics or 'forecast verification' of meteorological models where decisions regarding the capability of the model to represent the real-world must be made; and,

- Its wide range of applications, ranging from two-dimensional data fields from tests on engineering structures to three-dimensional data fields for which volumetric data are available relating spatially and temporally varying temperatures, where this methodology could be used to identify significant changes between measurements and predictions, or between successive measurements obtained over time indicating the change in condition of a system, such as the El Niño Southern Oscillation.

The proposed methodology supplements techniques for reducing the dimensionality of information-rich data fields by permitting the associated uncertainty in the data to be characterised and represented in the low-dimensional or feature vector space. The examples presented show that the results from the proposed methodology can be presented in a visual and intuitive manner to inform decision makers about the uncertainty in data and the significance of differences between data fields whilst allowing multivariate statistics to be utilised so that inferences can be drawn.

Chapter 5

Model validation using spatial measurements

5.1 Introduction

Computational models are used across scientific disciplines to make predictions and aid the decision-making process. In engineering, models are commonly employed to simulate the response of a structure for given operating conditions. However, computational models consist of numerical abstractions and regardless of the level of mathematical sophistication, it is important for engineers to possess confidence that these numerical surrogates can accurately represent the actual structure well enough that they can be exploited to draw meaningful inferences.

In order to test whether the simulations can accurately represent the real world, engineers go through a validation process which as defined by the ASME [4], is the process of determining the degree to which a model represents the real world from the perspective of the intended uses of the model. This process is reflected in figure 5.1 which demonstrates the triangle-shaped approach commonly employed by the aerospace industry during the development of new aircraft. The layers at each level of the triangle demonstrate the synergy between the modelling & testing activities across the different scales of development. This testing regime partly mandated by air-worthiness regulations and partly targeted by the companies' internal structures is aimed into gathering evidence that the under-

lying structure will be able to withstand the loadings under various conditions without endangering the safety of the passengers. It also offers confidence to the decision makers that the product of the design cycle responds to the initial objectives and meets the predefined quality requirements. This approach entails tests starting at the coupon level and gradually increase in complexity and scale to subcomponent, component and finally, full scale tests. The number of these tests can range from hundreds to thousands at the coupon and element level while it rapidly decreases when moving towards full scale testing.

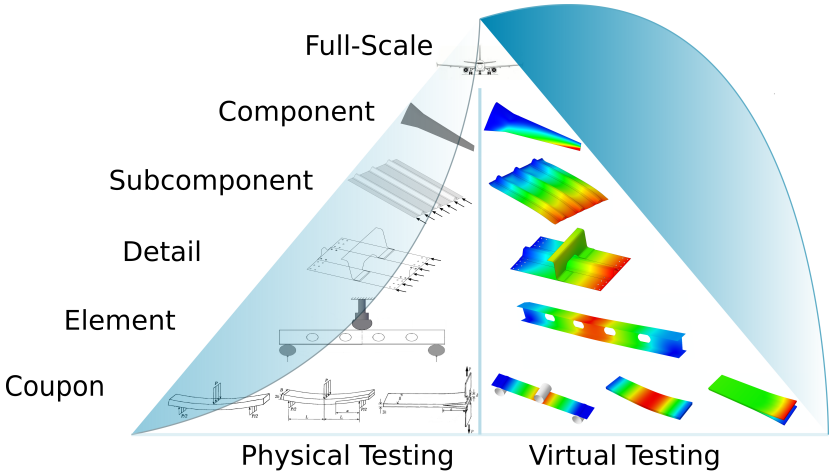

Figure 5.1: Triangle-shaped approach employed during the development of new aircraft.

With evolutionary rather revolutionary changes in aircraft design for the most part over the last three decades and with increasing levels of fidelity in terms of scientific theories and computational power one could suggest that it is time that physical testing could be replaced by virtual testing. This ambition is reflected by the colour intensity of the two arcs depicted in the same figure; one decaying in the left side, from bottom to top, demonstrating the aim to reduce costly and time-intensive tests at the higher scales and one increasing in the right side, reflecting the need for accurate, high-fidelity computational simulations. However, to be able to bring this aspiration to fruition, established validation techniques capable of quantitatively characterizing the accuracy of simulations across the different levels of testing should be in place.

For the case of structural mechanics models, a series of instructions in the form of a guide have been suggested by the CEN [8]. One of the novelties of

the guide lies in the use of full-field measuring devices such as digital image correlation to validate structural models. This is advantageous to traditional, point measuring devices, as measurements can be taken across the whole region of interest, devoid of gaps or hidden spots. However, the process of comparing measured quantities to model predictions is not that straightforward as the two datasets will probably lie on different grids. An elegant way to avoid this problem is by employing orthogonal decomposition techniques using Chebyshev or Zernike polynomials that allow the representation of the different datasets in the form of a vector through a number of commonly extracted features.

As shown in the previous chapter, these datasets can be infected with measurement uncertainty which can be spatially constant or varying and has to be taken into consideration during validation. In this chapter, two methods for the validation of computational structural models using full-field measurements will be shown. The first one is based on a pixel wise comparison of the predicted against the measured dataset and is achieved via the decomposition and subsequent reconstruction of the two datasets on the same grid. Afterwards, a probabilistic statement describing the percentage of differences lying within the bounds formed by the measurement uncertainty is made and conveyed to decision makers.

In the second method, the prediction and the measurement are initially transformed into their equivalent feature vectors. Afterwards, the measurement and its uncertainty are mapped into a probability distribution using the approximate Bayesian computation (ABC) technique described in the previous chapter. Finally, the Mahalanobis distance, a distance measure between a probability distribution, which in this scenario represents the measurement and its uncertainty, and a point in space, which now consists of the prediction's feature vector, is used to assess the quality of the model.

A series of measurements including displacements and deformations across two regions of interest in an aluminium I-Beam in 3-point-bend loading will be used to demonstrate the applicability of the developed techniques to real-world examples.

5.2 Methodology

The two aforementioned techniques developed for the validation of structural mechanics models using full-field measurements will hereby be described in detail.

5.2.1 Pixel-wise probabilistic metric

A feature extraction technique, namely orthogonal decomposition using Chebyshev polynomials is initially used to decompose the measured and predicted datasets into two vectors, each one containing a number of features. The number of features used to represent these datasets is carefully selected so when they are reconstructed most of the initial information is retained. In the case of structural mechanics, criteria established for the selection of the number of coefficients to be retained have been suggested by the CEN guide [8] as described in the previous chapter. Care should be taken so that both datasets are represented using the same features.

After the initial decomposition, the two datasets are reconstructed on the same grid which allows a pixel-wise comparison. In other words, their differences are calculated at each grid location and the percentage of those differences that lie within the interval defined by the measurement uncertainty is calculated and represents the outcome of the validation process. This percentage represents the proportion of predicted-measured differences that can be explained by the presence of measurement uncertainty.

Mathematically the probabilistic metric can be described as:

$$PM = \frac{1}{kl}\sum_{i=1}^{k}\sum_{j=1}^{l}\mathbb{1}\{|m(i,j) - s(i,j)| \leq 1.96u_{meas}(i,j)\} \tag{5.1}$$

where $m(i,j)$ and $s(i,j)$ correspond to the measured and simulated fields following their reconstruction on the same grid, while k, l is the number of pixels in the horizontal and vertical direction. The symbol $\mathbb{1}$ represents the indicator function while the value of 1.96 in the right hand of the above equations corresponds to the extent of the measurement uncertainty $u_{meas}(i,j)$ that can explain the pixel-wise differences. This equation can be considered an expansion of equation (4.2) that was used to characterize whether a synthetically generated dataset was

representative of the measurement and its uncertainty during the ABC process. Compared to that accept/reject output, equation (5.1) can be used to assign a percentage reflecting the similarity between the simulated and measured datasets given the uncertainty in the latter.

As reported in the literature review, the measurement uncertainty can be accurately represented probabilistically using the Gaussian distribution. The Gaussian or Normal distribution is fully characterized by the mean, μ, and the standard deviation, σ. The adoption of the Gaussian to represent the measurement uncertainty allows the establishment of 95% confidence intervals. This means that if the difference between two pixels is outside the interval defined by $[-1.96\sigma, 1.96\sigma]$ then is deemed to be statistically significant and cannot be justified be the presence of measurement uncertainty alone.

The mean μ can be estimated through the arithmetic mean of n repeated observations, while the randomness, related to the standard deviation, σ of the Gaussian distribution can be obtained from a calibration procedure and can be constant or spatially varying. In the examples shown in figure 5.2 the measurement uncertainty is spatially constant with a magnitude of 0.01 mm for displacements and 30 $\mu\epsilon$ for strain measurements. These magnitudes along with their spatial distribution are the result of a calibration exercise published in [9].

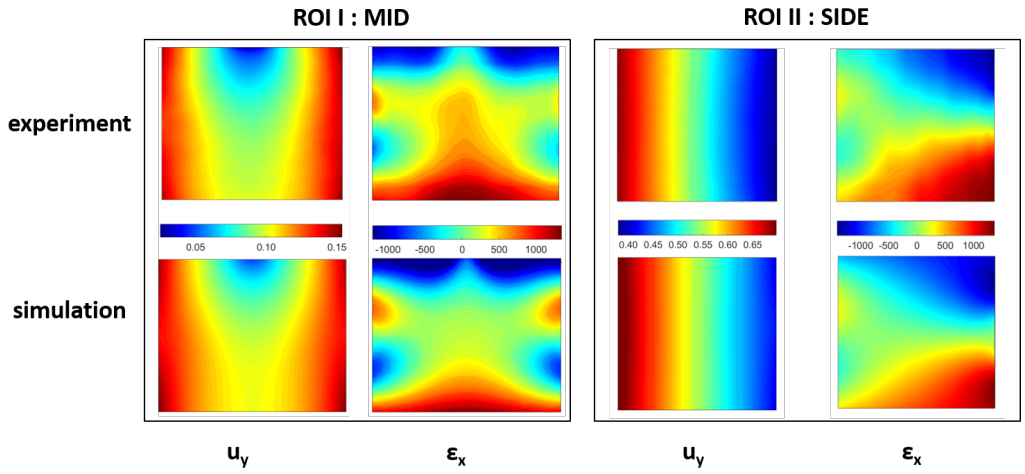

Figure 5.2: Datasets used for the demonstration of the proposed validation techniques. The u_y displacemements are in mm while the ϵ_χ deformations are in $\mu\epsilon$.

A series of engineering examples along with detailed visualizations will be provided in the following sections to aid the understanding of the developed

method.

5.2.2 Mahalanobis distance-based validation metric

The Mahalanobis distance [63] is hereby used to characterize the difference between the predicted and the measured data fields. In its general form, the Mahalanobis distance between a point $\mathbf{x}_i$ and the mean of a distribution $\overline{\mathbf{x}}$ is calculated as

$$MD_i = \sqrt{(\{\mathbf{x}_i\} - \{\overline{\mathbf{x}}\})^T [\mathbf{C}_x]^{-1} (\{\mathbf{x}_i\} - \{\overline{\mathbf{x}}\})^T} \tag{5.2}$$

where $[\mathbf{C}_x]$ is the sample covariance matrix of the latter calculated using equation (5.3) for the 2-D case. The covariance matrix, also known as the variance-covariance matrix, is a square matrix whose diagonal elements are the variances of the distribution in each dimension, while the off-diagonal elements are the covariances between variables/dimensions. The covariance between two variables X_i and X_j, is in its general form described by equation (5.4).

$$C_x = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{5.3}$$

$$C_{X_i X_j} = cov[X_i, X_j] = E[(X_i - E[X_i])(X_j - E[X_j])] \tag{5.4}$$

In the present validation setting both data fields are initially transformed into the corresponding feature vectors using the same orthogonal, Chebyshev polynomials as in the previous section. Then, the uncertainty-infected, measured data field is transformed into a probability distribution using the approximate Bayesian computation described in the previous chapter. This ensures that the measurement uncertainty, which is known in the spatial domain, is accurately represented in the feature vector domain.

Afterwards, in a way similar to the Mahalanobis distance area metric [66], the Mahalanobis distances of each of the sample points drawn during the ABC are calculated with respect to the distribution itself. The points near the mean of the distribution will produce values close to zero, while points away from the mean will result in larger values. This permits the formation of an upper bound

on what can be deemed an accurate representation of the measurement and its uncertainty as will be shown later.

Finally, the Mahalanobis distance of the predicted feature vector is calculated against the posterior distribution. This enables an assessment of the accuracy of the prediction in a quantitative manner while accounting for the measurement uncertainty. If the calculated distance is larger than the upper bound defined earlier, then the prediction is considered to be unrepresentative of the measurement. Otherwise, special care must be given to avoid mistakes as will be explained subsequently in detail.

This process can be mathematically described using the following steps:

1. calculation of the Mahalanobis distance (MD_{EXP_i}) for each of the sample points (x_{EXP_i}) drawn during the ABC with respect to the mean of the posterior ($\overline{x_{EXP}}$):

$$MD_{EXP_i} = \sqrt{(\{\mathbf{x_{EXP_i}}\} - \{\overline{\mathbf{x_{EXP}}}\})^T [\mathbf{C_{ABC}}]^{-1} (\{\mathbf{x_{EXP_i}}\} - \{\overline{\mathbf{x_{EXP}}}\})^T} \tag{5.5}$$

   Nominally, the mean of the posterior should be the feature vector of the measurement; C_{ABC} is the covariance matrix calculated from the samples that constitute the posterior distribution.

2. calculation of the Mahalanobis distance of the simulation output(s) with respect to the posterior distribution:

$$MD_{SIM_i} = \sqrt{(\{\mathbf{x_{SIM_i}}\} - \{\overline{\mathbf{x_{EXP}}}\})^T [\mathbf{C_{ABC}}]^{-1} (\{\mathbf{x_{SIM_i}}\} - \{\overline{\mathbf{x_{EXP}}}\})^T} \tag{5.6}$$

3. comparison between the value(s) of MD_{SIM_i} and the distribution of MD_{EXP_i}.

5.3 Applications

To demonstrate the applicability of the proposed methodologies for model validation using spatial measurements, a series of examples from aerospace engineering will be employed. The data were obtained by Lampeas et al. [9] from an aluminium I-beam in 3-point-bending using a set of stereoscopic cameras and a

digital image correlation system. The setup and the characteristics of the digital image correlation system are described in more detail in the previous chapter and in [9]. The measured and the predicted datasets including y-displacements and x-deformations are shown in figure 5.2. The measurements were obtained across two regions of interest (ROIs), ROI I and ROI II as shown in figure 5.3. The first region of interest (ROI I) is at the middle of the beam, right under the loading nose, while the second region of interest (ROI II) is between two holes, located at the left side of the web. Initially, the y-displacement data of ROI I will be used to provide a more-in-depth explanation of both methodologies, followed by a comparison of the quantitative results with other published techniques.

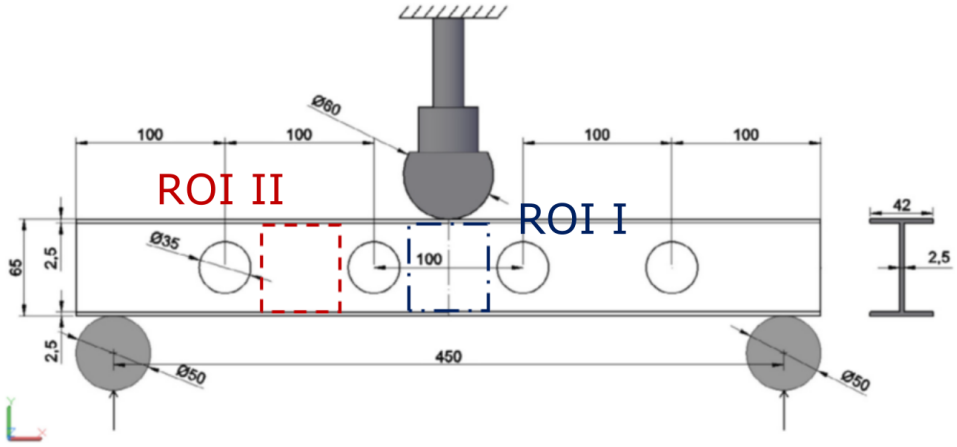

Figure 5.3: Schematic of the I-beam setup with the two regions of interest (ROIs) outlined (adapted from Lampeas et al. [9]).

5.3.1 Pixel-wise comparisons

The results of the pixel-wise assessment are shown in figure 5.4. It becomes apparent from the top-left corner of the figure that almost all of the u_y displacement differences are within the interval defined by the measurement uncertainty ($u_{meas} = 0.01$). To explain how the final percentage is calculated using equation (5.1) a graphical demonstration will be pursued with the aid of figure 5.5. It can be seen that the biggest differences are located on the top of the beam, at the contact point between the loading nose and the beam's web, as portrayed in more detail by the close-up image. Focusing on the single pixel pointed out by the grey arrow, it is evident that the magnitude of the differences is 0.024mm. This differ-

ence, visualized on the right side of the figure, as a vertical blue line, lies outside the 95% confidence interval shown, which suggests that the difference cannot be attributed to the measurement uncertainty alone. The orange curve represents the Gaussian distribution attributed to the measurement uncertainty, while the grey area along with the two vertical dashed curves represents the 95% confidence interval. The Gaussian is centered at zero implying that there is no bias in the measurement, while the standard deviation is equal to the measurement uncertainty, which as stated earlier is equal to 0.01 mm and spatially constant.

Repeating the same calculation at every pixel location across the grid using equation (5.1) produces the outcome of the process. In this case the result is 99.9% as almost all of the differences are within the range defined by the measurement uncertainty except the small region at the top.

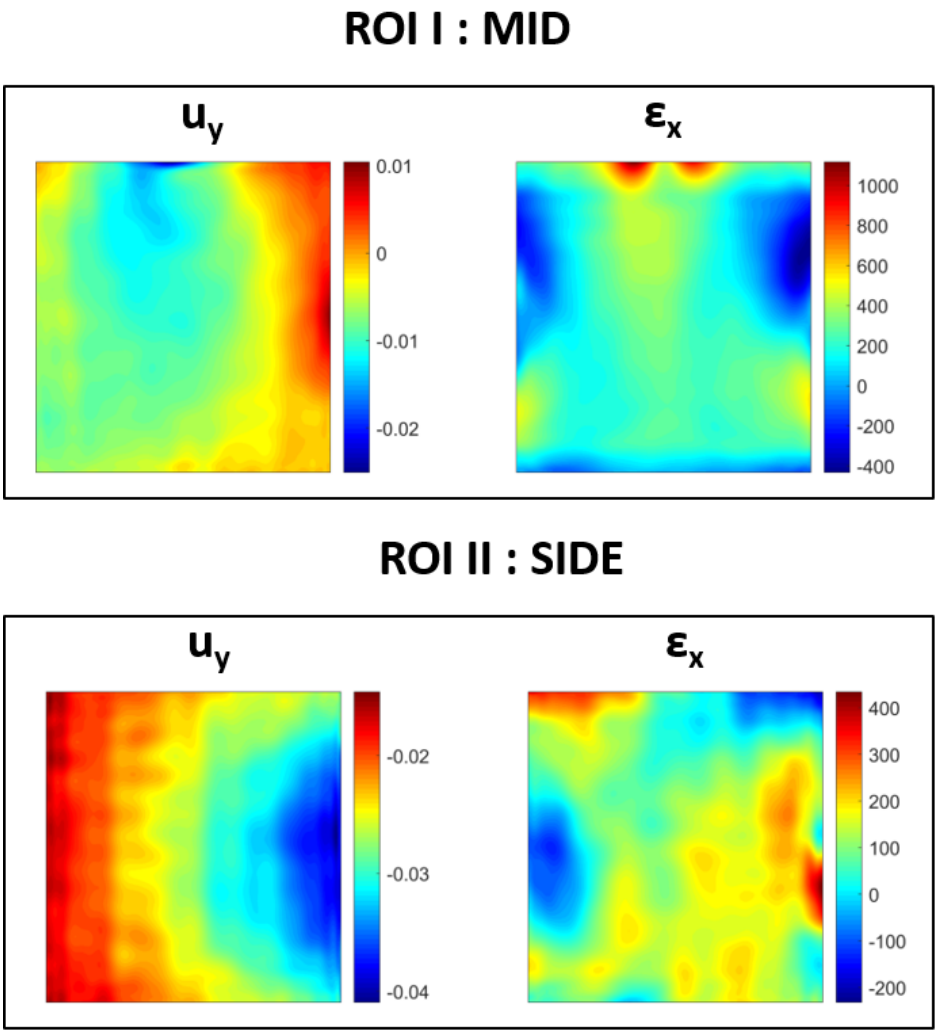

Figure 5.4: Pixel-wise differences for the two regions of interest. The displacement measurements (u_y) are in mm, while the deformation measurements (ϵ_χ) are in $\mu\epsilon$.

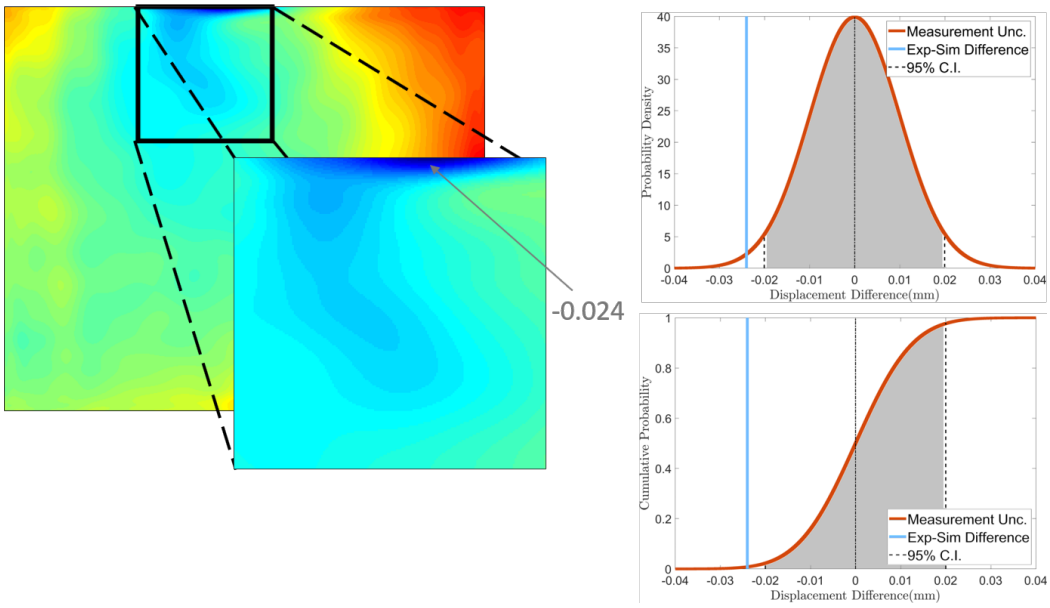

Figure 5.5: Close up view of the pixel differences for the u_y displacements at ROI I, along with a depiction of the statistical testing of differences given the presence of normally distributed measurement uncertainty, in the form of a probability density (top) and cumulative probability (bottom).

Even though visualizations, such as those in figure 5.4, can assist in identifying areas with the biggest differences, they fail in providing a quantitative overview of the data. An alternative, perhaps more intuitive way to achieve this is shown in figure 5.6 for the ϵ_χ strain differences at ROI I. The empirical distribution functions of the measured and simulated data are plotted against each other on the left side of the figure while on the right side the empirical distribution function of the pixel-wise differences is shown. It can be seen that only a small proportion of the differences is within the bounds defined by the measurement uncertainty, resulting in a value of 11.3%. This value can be directly calculated from the right graph of figure 5.6 as the vertical distance between the points where the empirical distribution intersects the two dashed lines representing the measurement uncertainty. Repeating the procedure for the remaining datasets produces the graphs in figure 5.7 while the outputs from the comparisons are shown in table 5.1.

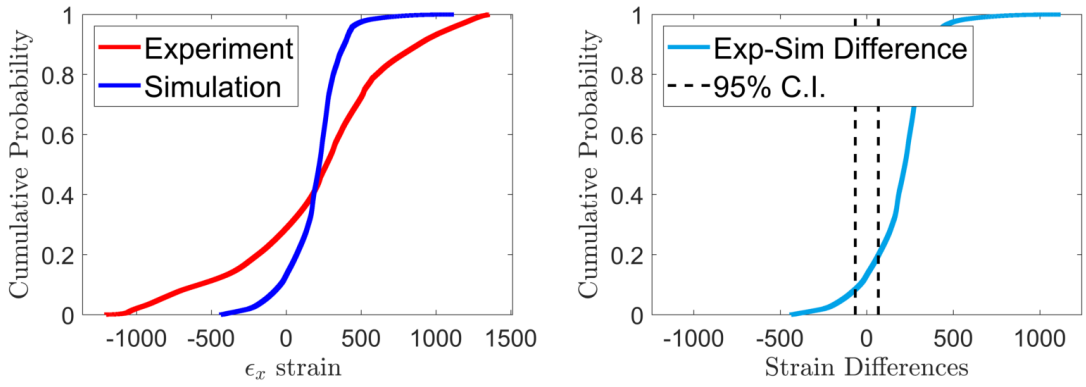

Figure 5.6: The empirical distribution functions of the measured and simulated fields for the ϵ_χ deformations at ROI I are shown on the left side, while on the right side the empirical distribution of the differences along with two vertical dashed-lines representing the expanded measurement uncertainty are given. The units are in $\mu\epsilon$.

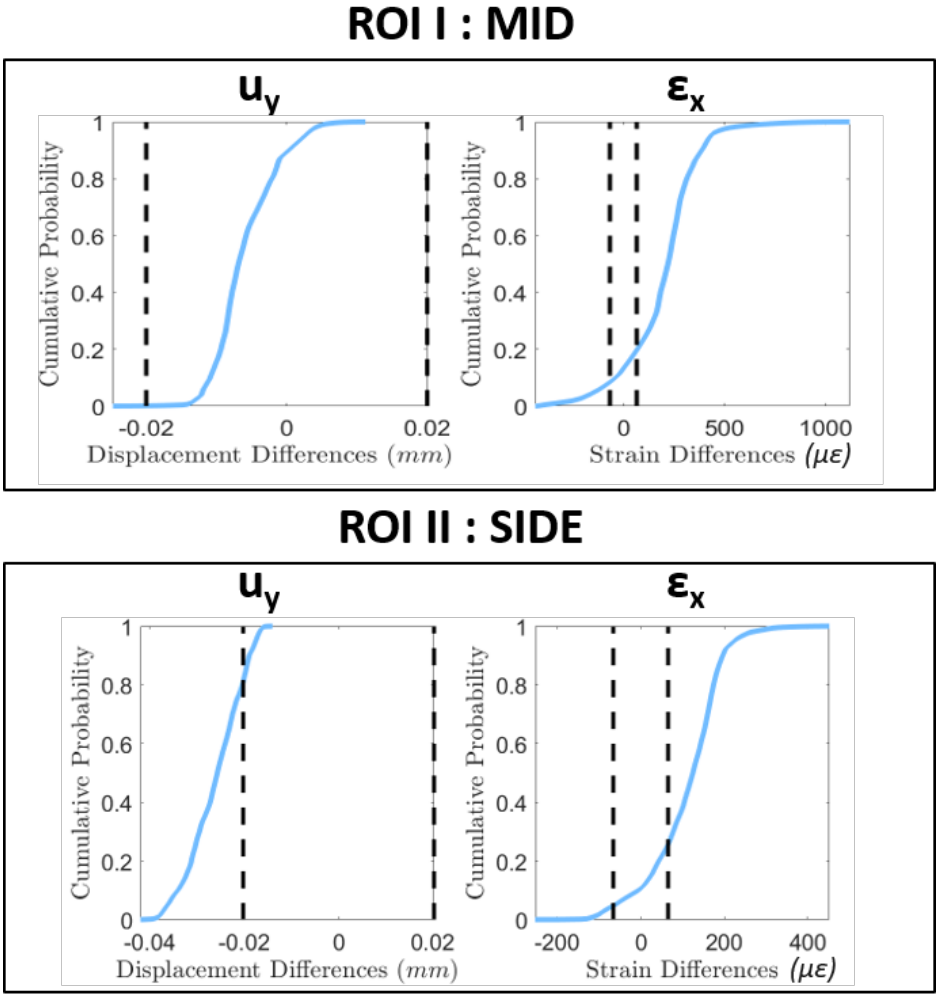

Figure 5.7: Pixel-wise differences plotted for the datasets portrayed in figure 5.2 as empirical distribution functions along with two vertical dashed lines representing the expanded measurement uncertainty.

Table 5.1: Probabilistic metric results for the data of figure 5.2.

ROI	Dataset	Probabilistic metric.
1	u_y	99.90%
1	ϵ_χ	11.33%
2	u_y	18.77%
2	ϵ_χ	20.82%

The conclusions that can be drawn from the findings in table 5.1 regarding the datasets portrayed in figures 5.4 and 5.7 are the following: for the predictions of the u_y displacements of ROI I, 99.9% of the pixel-wise differences lie within the band defined by the measurement uncertainty (1.96 u_E). This value determines the capacity of the model to accurately simulate the structure at that region for the given response. On the other hand, its accuracy in predicting deformations (ϵ_χ) both in ROI I and II is significantly less. Only 11% of the differences can be attributed to the presence of measurement error for the case of ROI I and a similar value (18%) is also observed for ROI II. Such low values cannot be attributed to the presence of measurement error alone, indicating that the reasons behind such results should be examined further. These could range from the way the loading is applied, to the boundary conditions, geometry and material parameters of the specimen or numerical post-processing. In a similar manner it is obvious that the predictions for the u_y displacements of ROI II deviate significantly from their measured counterparts. Even though the two fields may qualitatively look similar, these differences could be attributed to inaccurate stiffness parameter selection as the predicted field is obviously more compliant. A method to calibrate the model's parameters for this case has already been shown in figure 4.4.

5.3.2 Mahalanobis distance-based assessments

The second technique is based on the Mahalanobis distance to characterize the similarity between the predicted and the measured data fields. The Mahalanobis distance (MD) in its one-dimensional form is a measure of distance, in standard deviations, of a point from the mean of a distribution. The characteristic of the MD which is beneficial in the current validation context is that it is an extension

of the Euclidean distance normalized by the covariance matrix. This means that the distance calculation between the feature vectors of the two datasets accounts for the measurement uncertainty. Loci that are equidistant from the mean of the distribution against which the Mahalanobis distance is calculated, take the form of hyperellipsoids in high dimensional spaces when they are visualised. This is different from the commonly used Euclidean distance where these take the form of concentric hyperspheres. To provide a better explanation of this phenomenon, the u_y displacement data from ROI I will be employed. As described in the previous chapter, a total of 9 Chebyshev shape descriptors were used to accurately represent the dataset in its low-dimensional form following the CEN recommendations [8]. Given that visualizations depicting more than three dimensions can be quite troublesome to interpret, a 2-D visual explanation subsequently followed by its 3-D expansion, of the Mahalanobis distance will be provided.

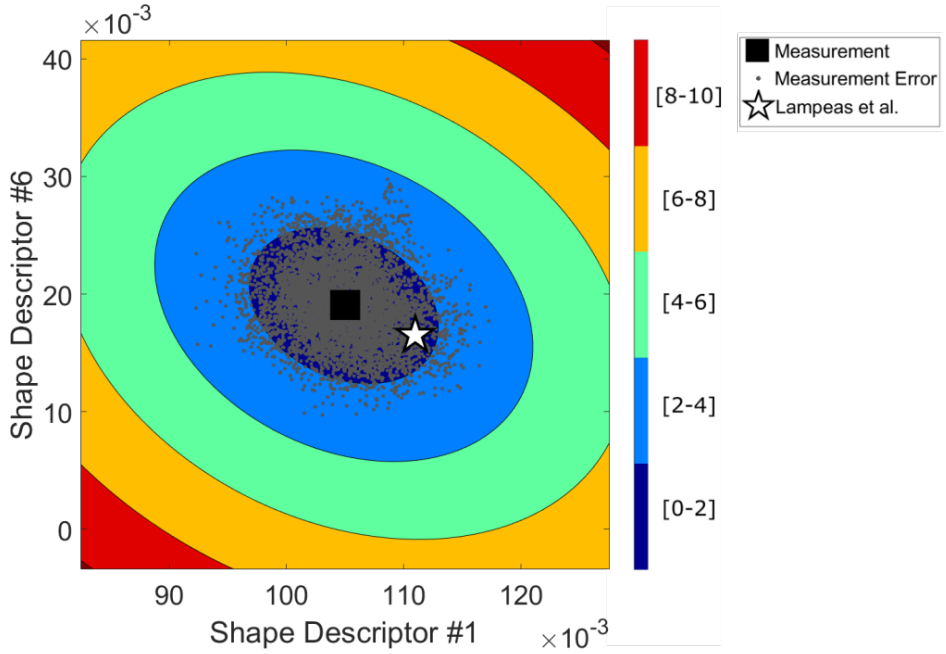

Figure 5.8: The measurement and its uncertainty for the u_y, ROI I, displacement data are shown as a black square and surrounding grey circles respectively, while coloured contours depict the Mahalanobis Distance. The formation of ellipses reflects the equidistant loci in this 2-D example.

Retaining the notation of figure 4.6, the u_y displacement, ROI I, measurement along with its uncertainty and the prediction made by Lampeas et al. [9] for the first and sixth shape descriptors are shown in figure 5.8. The various ellipses centered around the measurement reflect the covariance structure of the posterior

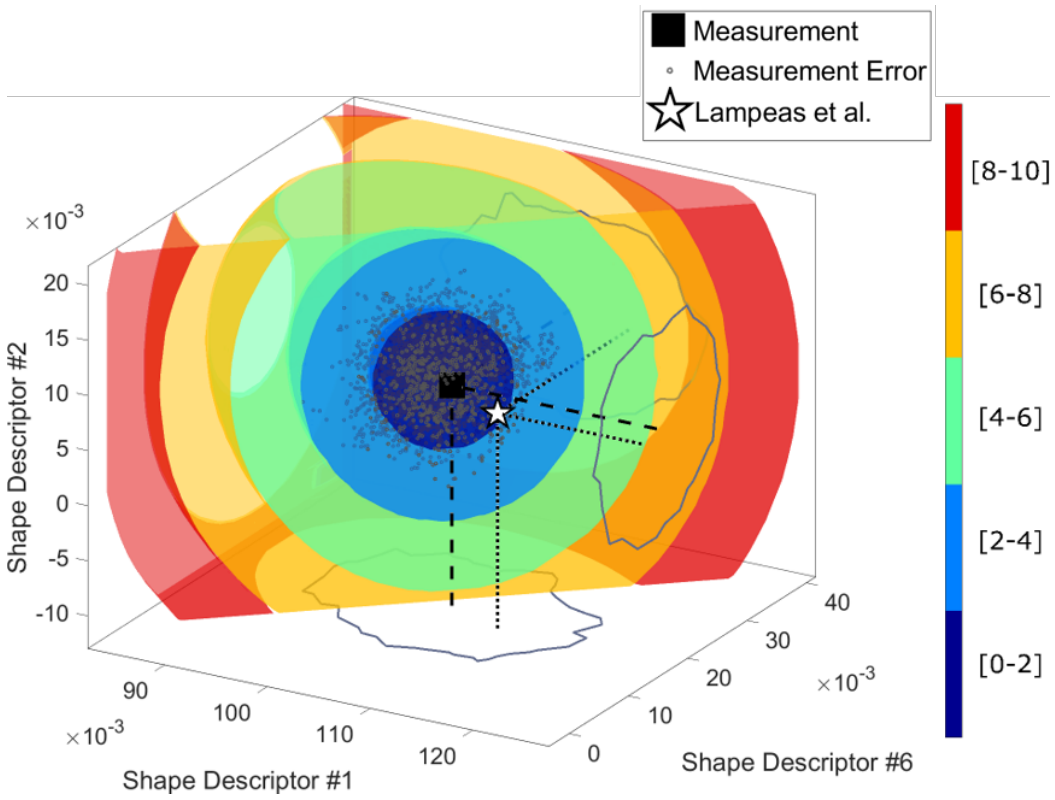

Figure 5.9: 3-D extension of the Mahalanobis visualization of figure 5.8 for the u_y displacement data.

distribution, while the colouring represents the range of Mahalanobis distances with respect to the mean of the distribution within each region as shown by the colourbar. It can be seen that for the 2-D case, the Mahalanobis distances of the points corresponding to the measurement and its uncertainty are less than 4, while the Mahalanobis distance of the prediction made by Lampeas et al. [9] is less than 2. In this 2-D simplification this means that Lampeas' prediction accurately represents the real world as it is closer to the measurement than some of the more distant samples that constitute the distribution.

Extending the visualization of the Mahalanobis distance to three dimensions results in figure 5.9. This figure is an enhancement of figure 4.6 with hyperellipsoids drawn to visualize the Mahalanobis distance. The emerging correlation between shape descriptors #1 and #6 becomes apparent by the orientation of the hyperellipsoids in the $x-y$ plane. Moreover, a closer look on the grey curves outlining the posterior distribution across the three planes suffices to conclude that the formed hyperellipsoids may not precisely fit that geometry. This issue and its implications on validation will be discussed later in the chapter.

As stated earlier, visualizing high-dimensional quantities such as the feature vectors representing the measured and simulated datasets is not straightforward. A simple way to circumvent this problem, while accurately depicting the probabilistic nature of the measurement and its uncertainty, is by plotting the data through a series of 2-D histograms each one representing certain feature combination as demonstrated in figure 5.10 for the same u_y displacement measurement. The rows have been arranged in descending order in that the top row corresponds to the shape descriptor with the biggest magnitude which in this case is the first and the rest follow in sequence. The same ordering has been applied to the columns from left to right, while the shapes of the corresponding features have been plotted on the main diagonal. As an example, the top left shape corresponds to the first shape descriptor which is the average of the measured quantities, the second shape of the diagonal corresponds to the sixth descriptor and so on. Each graph is centered on the measurement for the corresponding feature combination while the colouring reflects the probability density at the given bin location. It is obvious that the probability density is maximized around the measurement, as demonstrated by the dark red colour at the centre of each subfigure. The prediction by Lampeas et al. [9] has been overlaid using a brown triangle, which allows for a quick visual, qualitative evaluation of the distance between the prediction and the measurement. It can be easily observed that the prediction lies inside the domain defined by the posterior distribution in the 9-D space, while the largest deviation occurs in the shape descriptor #1. The Mahalanobis distance between the simulation and the measurement is in this case 1.76.

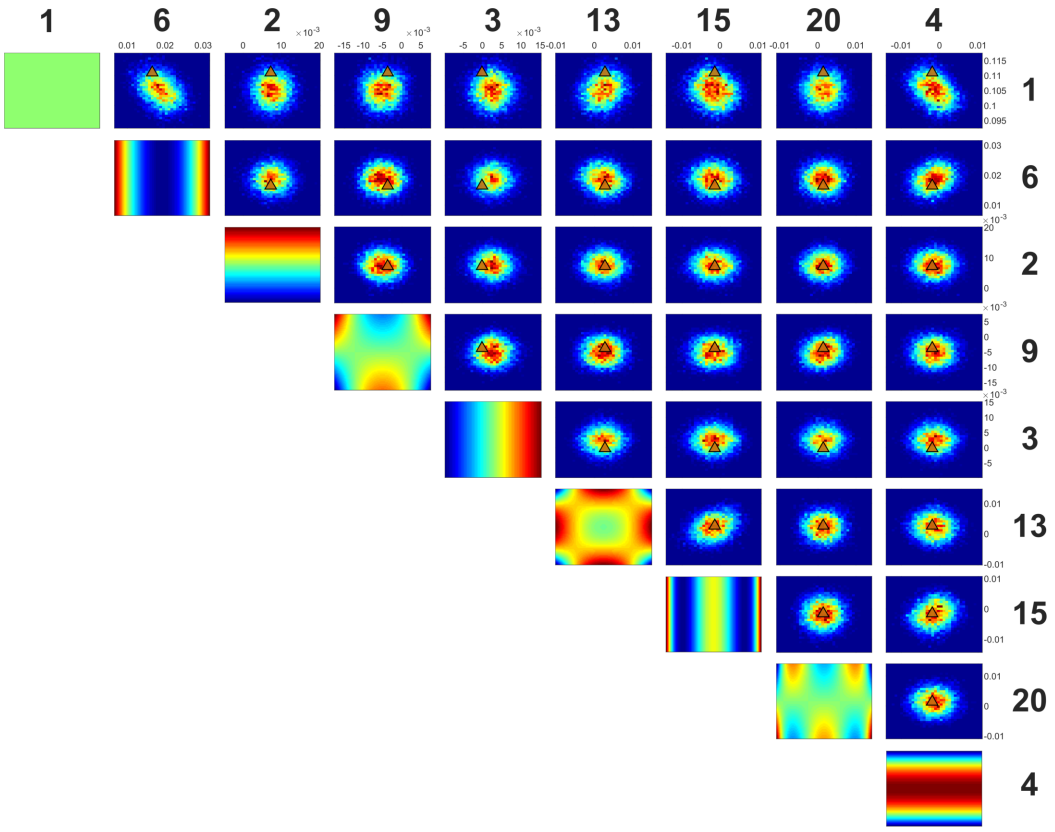

Figure 5.10: A series of 2-D histograms visualizing the measurement uncertainty for each shape descriptor combination for the u_y displacement data of ROI I as probability densities. The shapes at the main diagonal correspond to the shapes of the respective descriptors. The prediction is shown as a brown triangle in each graph.

A counterexample where the prediction is far from the measurement is given in figure 5.11 for the ϵ_χ deformations in ROI II. Due to the complexity of the deformation field shown in figure 5.2 a total of 96 shape descriptors were used for the decomposition, nine out of which have been plotted due to space limitations. It is apparent that the largest discrepancies are in the eighth and first shape descriptors, while the magnitude of the measurement uncertainty is reflected by the small area at the centre of each figure thus leading to a Mahalanobis distance of 87.38 between the two datasets. The results for the remaining datasets along with the number of shape descriptors used can be seen in table 5.2.

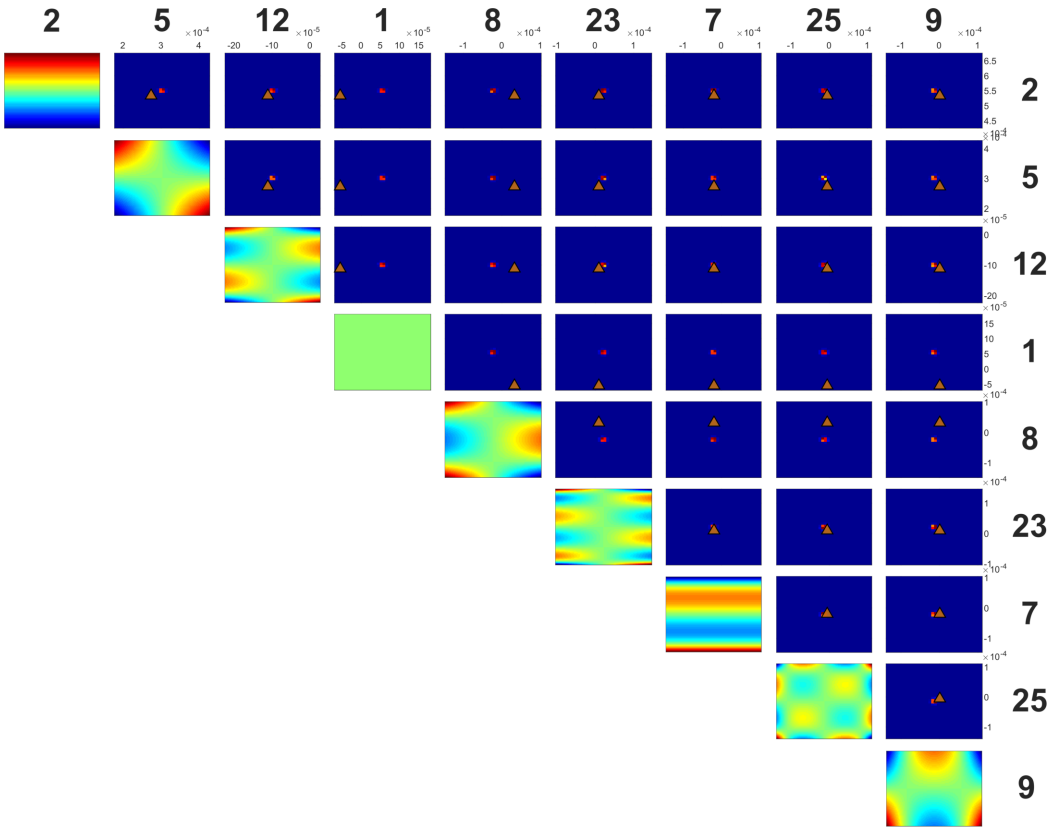

Figure 5.11: 2-D histograms illustrating the measurement and its uncertainty for the ϵ_χ, ROI II dataset. The lack of overlap between the measured and the predicted feature vectors implies that the prediction does not accurately represent the measured quantities.

Table 5.2: Mahalanobis distance calculation along with upper bounds for the data of figure 5.2.

ROI	Dataset	No. shape descriptors	Mahalanobis distance	Max. Mahalanobis
1	u_y	9	1.76	4.22
1	ϵ_χ	96	87.38	12.11
2	u_y	2	3.29	2.42
2	ϵ_χ	69	35.56	10.89

5.4 Discussion

5.4.1 Determination of upper bound for the Mahalanobis distance

Even though the Mahalanobis distance can be used to quantitatively assess the similarity between a simulation outcome and a measurement in the feature vector domain, yet it fails in describing intuitively the extent of that similarity. Unless one is faced with a trivial situation like one in which the measurement and its uncertainty is represented by a univariate normal distribution and a simulation that is one standard deviation away from the mean of the normal, it is difficult to intuitively understand how similar is the simulation compared to the measurement. This issue is exacerbated in higher dimensions as the number of emerging correlations across the shape descriptors is increased while intuitive comprehension of the Mahalanobis distance is deteriorated.

For the case of the I-beam, the Mahalanobis distances between the simulations and the measurements by Lampeas et al. are outlined in table 5.2 along with the number of shape descriptors used to characterise the two fields in each case. Even though the Mahalanobis distance provides a quantitative method to compare measurements to simulation outputs, it makes it impossible to assess whether these simulations accurately represent reality. What is missing is an upper bound that would allow decision-makers to identify whether the simulation (or simulation ensemble) is within the range of the distribution that characterizes the measurement and its uncertainty without having to resort to visual aids. To determine this bound, a modification to the method proposed by Zhao et al. [66] for the validation of probabilistic models has been implemented. Initially, the Mahalanobis distance of each of the samples drawn during the ABC is calculated with respect to the mean of the posterior distribution. This step results in a distribution of Mahalanobis distances whose upper bound is defined by the most distant sample. Afterwards, the Mahalanobis distance of the simulation's feature vector is calculated with respect to the posterior distribution and compared to this bound. If the resulting distance is larger than the upper bound, then the simulation is determined to be unrepresentative of the measurement and its

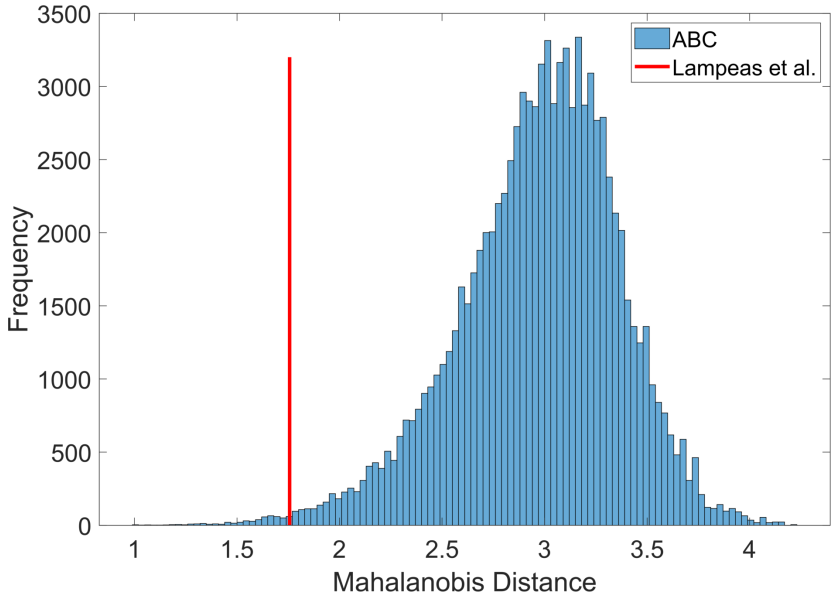

Figure 5.12: Distribution of Mahalanobis distances corresponding to the samples drawn from the posterior during the ABC away from the mean of the posterior. The red line corresponds to the Mahalanobis distance of the simulation by Lampeas et al.

uncertainty.

An example of this process can be seen in figure 5.12 for the case of the u_y displacement data of ROI I whose feature form is depicted in figure 5.10. The distribution of the Mahalanobis distances of the samples drawn during the ABC is shown in a histogram in which the maximum Mahalanobis distance of 4.22 acts as the upper bound on what could be considered representative of the measurement and its uncertainty. The Mahalanobis distance corresponding to the simulation made by Lampeas et al. has been overlaid using a red vertical line. Their prediction lies at the 0.5 percentile of the ABC Mahalanobis-transformed samples. The percentiles for the rest of the cases are not included as they lie outside the range of acceptable values. It is obvious that in this case the simulation's distance is within the range of acceptable values thus deeming it representative of the measurement. This result confirms the earlier qualitative evaluation made using figure 5.10 .

Worden and his colleagues [107] proposed a similar method for model validation based on earlier work [105] in the field of damage detection. Their aim in [105] was to identify the onset of damage using features extracted from the response of an operating structure. To do so they developed a technique that enables the determination of confidence intervals for the maximum allowable

squared-Mahalanobis distance based on a series of measurements corresponding to normal operating conditions. This practically means that the problem of damage detection was transformed into a hypothesis testing one and could be transferred into the context of model validation, where the normal operating conditions represented by the measurements are compared against model predictions. Their approach assumes that the measurement error infecting their experimental measurements, is normally distributed. This assumption, which allows the determination of confidence intervals through Monte Carlo sampling, could be a hindrance in cases where normality cannot be rightfully assured. For example, for the case of approximate Bayesian computation that is used to represent the measurement and its uncertainty in the feature vector domain, it becomes apparent from figures 4.5 and 4.7 that the posterior distribution is not Gaussian and making such an assumption would be unreasonable. Figueiredo et al. [155] modified this approach for the determination of the threshold. Instead of taking Monte Carlo samples from an assumed multivariate Gaussian distribution they set that threshold equal to the maximum Mahalanobis-squared distance from the feature-extracted measurements corresponding to the undamaged condition. Their approach is similar to the one taken here, where instead of having multiple measurements in the undamaged state, multiple samples taken from the posterior distribution represent a single measurement and its uncertainty in the feature vector space.

Caveats

The process described above outlines a way of characterizing whether a model prediction is acceptable, given its Mahalanobis distance from the distribution corresponding to the measurement and its uncertainty. Even though the whole process may seem straightforward, special care should be given on certain parts to make sure that errors are avoided.

The difference in the behaviour of the Mahalanobis distance and the pixel-wise probabilistic metric, with the latter being inherently employed during the ABC described by equation (4.2), may result in inconsistencies in what can be considered an acceptable representation of the measurement in the feature vector

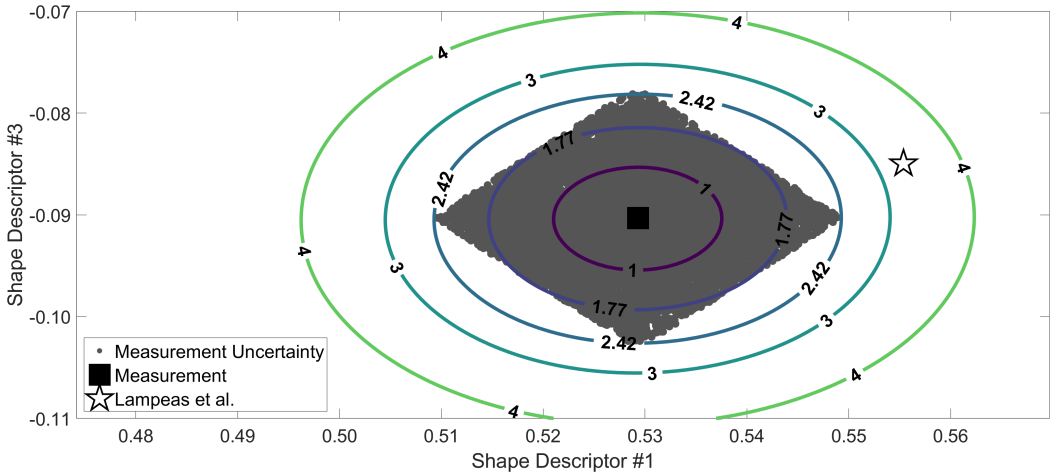

Figure 5.13: Isocurves outlining regions of equidistant Mahalanobis distances for the u_y displacement data of ROI II. The rhombus-shaped locus at the center reflects the samples drawn from the posterior distribution during the ABC. The prediction made by Lampeas et al. has been overlaid using a white star.

space. Analytically, the ABC results in a posterior distribution representing the measurement in the feature vector space which is defined over a certain locus, abiding to the constraint of equation (4.2). However, the information about the shape of this locus is partially lost when the Mahalanobis distance is used. The Mahalanobis distance is a distance metric solely characterised by the mean and the covariance structure of a distribution with respect to a point. This results in ellipse-shaped loci, examples of which are shown in figures 5.8 and 5.9 and which will usually not resemble the locus over which the posterior distribution is defined unless the latter is a Gaussian distribution.

The linearly-varying u_y displacement data of ROI 2 will be used to explain this phenomenon in more detail in figure 5.13. The advantage stemming from the linearity of the data is that they can be accurately characterised using only two shape descriptors and thus be plotted in the two-dimensional space. A series of colored ellipses corresponding to isocurves of Mahalanobis distances centered around the measurement will be used to describe three different scenaria corresponding to different ranges of the Mahalanobis distance. The first scenario consists of points in this 2-dimensional space whose distance is larger than 2.42 which belongs to the most-distant sample drawn during the ABC. It can be seen from the figure that points in this range $(2.42, \infty)$ do not intersect the locus of the posterior distribution thus safely concluding that they are not representat-

ive of the measurement and its uncertainty. The second scenario corresponds to points whose MD is less than 1.77 as depicted by the ellipse inscribed to the diamond-shaped locus. It can be stated that simulations lying inside this set are representative of the measurement. Finally, special attention should be given to points whose Mahalanobis distance is in the range of $(1.77, 2.42)$. As it can be seen in the figure there are regions in this range which are part of the posterior distribution and regions that are not. This result stems from the different manner by which the Mahalanobis distance and the measure characterizing the 'acceptability' of a feature vector during the ABC are defined. To avoid confusion in this range when visual aids are not available, practitioners are advised to use the probabilistic measure described by equation (5.1) to certify whether a simulation is inside the locus of the posterior distribution.

To demonstrate an example of the last statement, the u_y displacement measurement of ROI I is used along with a series of Monte Carlo simulations where the Young's modulus of the beam varies between 65 and 75 GPa resulting in figure 5.14. The histograms corresponding to the Mahalanobis distances of the simulation outputs and the measurement and its uncertainty are depicted in the left-side of figure 5.15 in red and blue respectively. It is obvious in this case that the Mahalanobis distances of the simulation outputs are less than the maximum Mahalanobis distance attributed to the most distant sample drawn during the ABC. This finding could lead to the false conclusion that all of the simulation outputs are representative of the measurement. However, a closer analysis with the aid of the probabilistic metric shown in the right side of the same figure leads to the fact that some of the predictions would be wrongly classified as representative of the measurement. It can be seen that in three of the total 72 simulation results, their probabilistic assessment is less than 95% which has been selected as the threshold value that characterises the acceptability of a simulated response compared to a measurement and its uncertainty.

5.4.2 Comparison of the proposed metrics

It becomes apparent from figure 5.15 that even though the two metrics, inherently employ equation (4.2) to assess the similarity between spatial fields, they

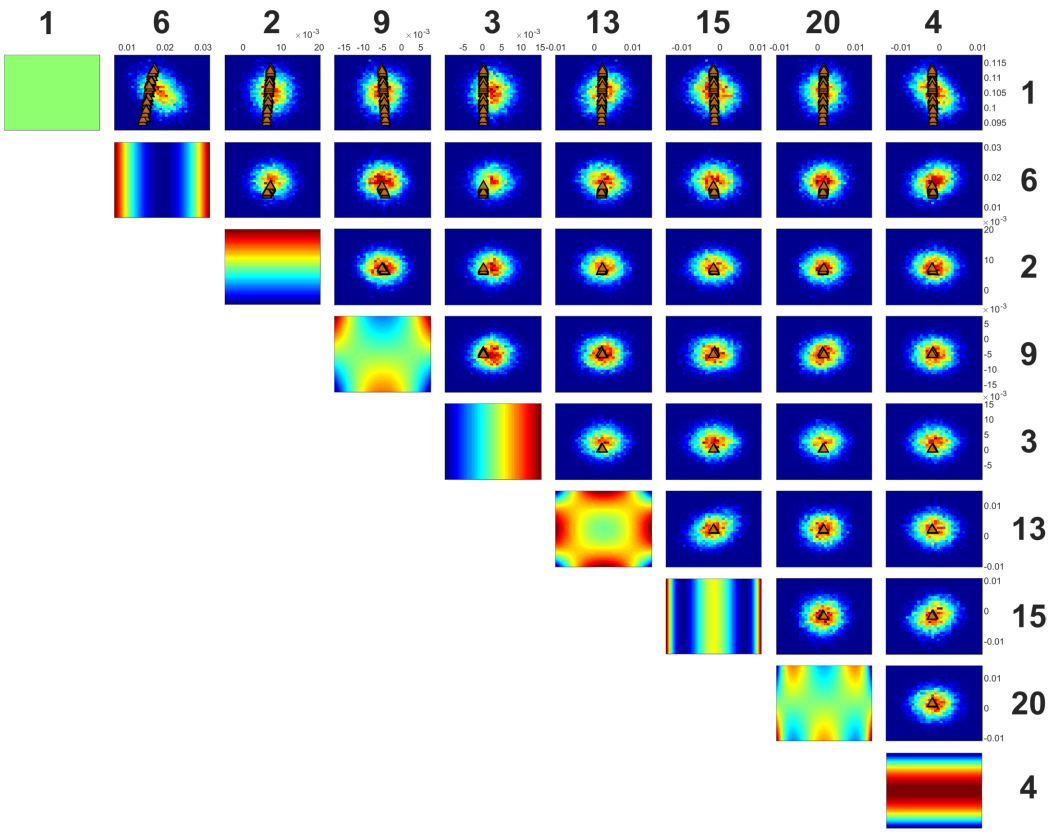

Figure 5.14: The results of a Monte Carlo simulation where the stiffness of the beam varied between 65 and 75 GPa for the u_y displacement data corresponding to ROI 1 are shown in brown triangles. The measurement and its uncertainty are depicted using 2-D histograms.

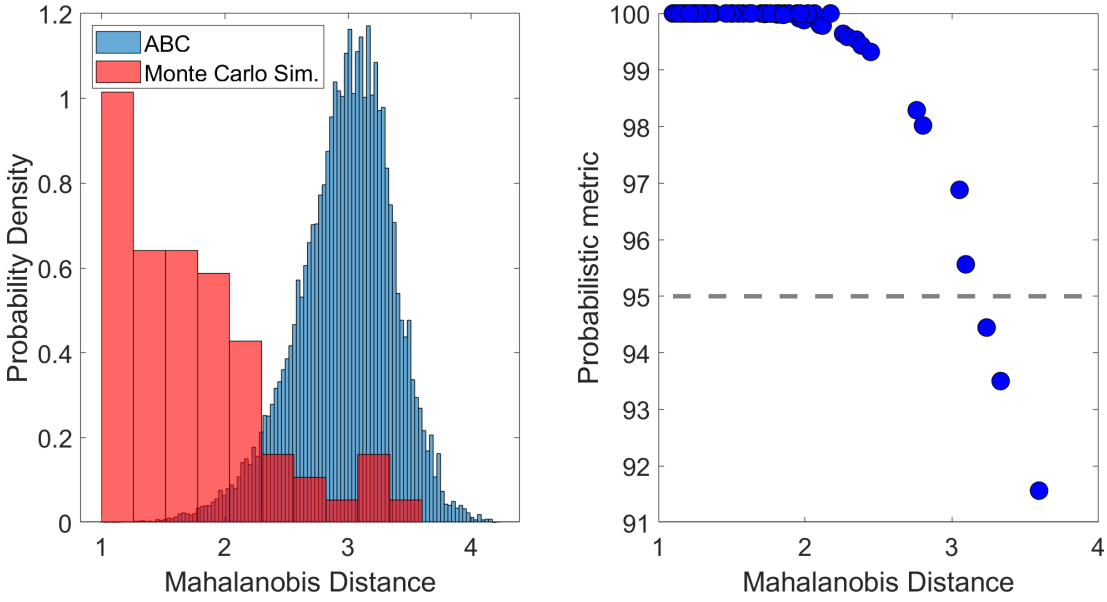

Figure 5.15: The distribution of Mahalanobis distances corresponding to the samples drawn from the posterior during the ABC are shown in blue. The red histogram corresponds to the Monte Carlo simulations of figure 5.14.

demonstrate some differences. These differences, along with the advantages and disadvantages of each metric will be discussed in more detail in this section. This will aid the reader identify which one to use for their application.

The prevalent similarity between the two is that both employ orthogonal decomposition to homogenize data potentially lying on different grids: the probabilistic metric to ascertain that the reconstructed simulated and measured fields are in the same grid following their initial decomposition, allowing for pixel-wise comparisons and the Mahalanobis distance to calculate their similarity based on their feature vectors. Even though the decomposition process is fundamental to both techniques its inherent cost in the case of the Mahalanobis distance can be considered higher. Analytically, in order to accurately represent the measurement uncertainty in the feature vector space during the ABC, the measured field is repeatedly assessed against synthetically generated ones to identify whether these could be considered representative of the former. However, this iterative approach can be time-intensive (similar to any Markov chain Monte Carlo technique) as a big number of evaluations are needed to accurately construct the posterior distribution. Compared to the thousands of evaluations needed to achieve this, the probabilistic metric proves to be a much more efficient alternative where the measured and simulated fields have to be reconstructed on the same grid only once.

Another benefit stemming from the use of the probabilistic metric is the capacity to spatially visualize the pixel-wise differences as in figure 5.4. This visual assessment can provide important information on the location and magnitude of the differences between the two fields. On the other hand, the Mahalanobis distance and visualizations such the one of figure 5.15 can be employed to identify the largest deviations across shape descriptors, while providing a simple way to depict the measurement uncertainty in the same space.

The capacity of the Mahalanobis distance to accurately characterise whether a simulation is representative of the measurement or not is adversely affected when the distribution that represents the measured field deviates from normality as is evident in figure 5.15. On the other hand, the fact that the Mahalanobis distance is unbounded as a distance metric provides a clear advantage to the probabilistic

metric that is bounded between 0 % and 100 %. This difference could prove to be vital during numerical processes (e.g. model updating or optimisation) where series of 0 % values are less informative compared to gradually decaying Mahalanobis distances as the algorithm searches towards the optimum.

Finally, it should be mentioned that the outcome of the probabilistic metric is simpler to decode, given that it represents a percentage and thus be easier to adopt in during decision making. On the other hand, the Mahalanobis distance, even though quantitative could prove hard to objectively assess, and would be more likely used to identify the best among competing models.

5.4.3 Comparison of the validation methodologies with other published techniques

As the rapid development of technology and science have resulted in unprecedented levels of quality measurements and detailed simulations, so has the demand for methods capable of quantifying the accuracy of these predictions. In structural mechanics a first step towards a standardised approach to validating computational models was taken through the CEN guide [8]. The members of the committee presented a method of characterizing the accuracy of a structural model using full-field measurements. The advantages were multiple as the newly-developed guide was the first outlining a series of practical recommendations required for the successful implementation of a validation procedure. One of those was the requirement for the acquisition of full-field measurements and the comparison of simulations and measurements in terms of their feature vectors.

The feature-based comparison suggested by the CEN is demonstrated graphically on the left-hand side of figure 5.16 for the u_y displacement data of ROI I. In this figure the measured shape descriptors are plotted against the predicted ones. The dashed line corresponds to the ideal scenario where the corresponding shape descriptors are equal, thus having a gradient of one, while the two adjacent continuous lines reflect the expanded measurement uncertainty $\pm 1.96 u_E$. When the points lie inside the aforementioned band then the prediction is deemed acceptable, while, on occasions that there are points outside the band the model is

rejected as described by equation (5.7).

$$\{\mathbf{s}_M\} = \{\mathbf{s}_E\} \pm 1.96 u_E \tag{5.7}$$

where $\{\mathbf{s}_M\}$ and $\{\mathbf{s}_E\}$ correspond to the shape descriptors from the model and the experiment respectively.

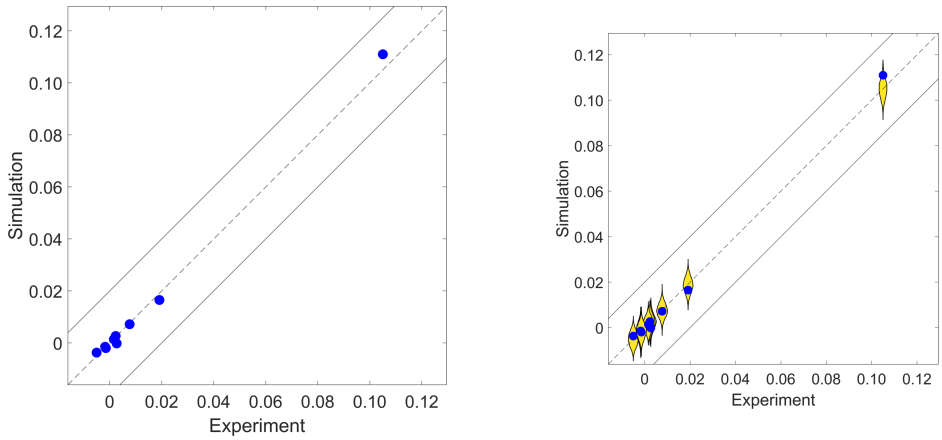

Figure 5.16: CEN [8] suggested plots for the validation of solid mechanics models. The 45°dashed line represents the ideal scenario where $\{\mathbf{s}_M\} = \{\mathbf{s}_E\}$, while the adjacent continuous lines reflect the expanded uncertainty defined by equation (5.7). On the right side of the figure, yellow violin plots are used to demonstrate the probabilistic nature of the measurement uncertainty in the feature vector space resulting from the ABC.

Even though this graph provides a simple way for a decision maker to identify whether their prediction represents the real structure, yet it misrepresents the extent of the uncertainty in each feature. One way of improving the existing figure while accurately depicting the uncertainty in the measurements is demonstrated on the right side of figure 5.16 for the case of spatially constant measurement uncertainty. The overlaid yellow symbols are known as violin plots and each symbol represents the marginal probability density of each descriptor as the result of the approximate Bayesian computation outlined in the previous chapter. They have all been positioned vertically on the 45°line and their width at each point represents the probability density reflecting the measurement and its uncertainty for the given ordinate. This means that as the distance between the measurement and a simulation increases, the lower is the probability of the latter to be representative of the former. It should be pointed out that even though it is the measurements that are infected with uncertainty, the decision to spread the

Table 5.3: Comparison of model validation results for the data of figure 5.2.

ROI	Dataset	Mahal. dist.	Max. Mahal. dist.	Probabilistic met.	Dvurecenska et al.	Lampeas et al.
1	u_y	1.76	4.22	99.90%	100%	Accept
1	ϵ_χ	87.38	12.11	11.33%	48%	Reject
2	u_y	3.29	2.42	18.77%	N/A	Reject
2	ϵ_χ	35.56	10.89	20.82%	100%	Accept

violin plots vertically was driven by the need to reduce the visual clutter added by the probabilistic representation of the measurement uncertainty, especially for the large number of coefficients whose magnitude is almost zero and the ease it allows for visual comparisons with the simulation's shape descriptors.

The validation outcomes for both of the proposed measures regarding the data depicted in figure 5.2 have been reported in table 5.3 along with the outcomes by Lampeas et al. [9] and Dvurecenska et al. [11]. It should be noted that the final outcome (accept/reject) in the results by Lampeas et al. is based on equation (5.7) and the employment of Zernike polynomials to decompose the datasets. On the other hand Dvurecenska et al. developed a relative error metric which employs orthogonal decomposition using Zernike or Chebyshev polynomials and accounts for the measurement uncertainty via a derived threshold.

The proposed techniques are in agreement with the outcomes by Lampeas et al. except the deformation field of ROI 2, where both Lampeas et al. and Dvurecenska et al. agree that the prediction represents the real world; the latter suggesting a 100% probability. Given the magnitude of local differences shown in figure 5.4 and the outcome of the probabilistic metric graphically shown in figure 5.7 it seems unlikely that the comparison of two datasets, where the absolute differences are more than 150 $\mu\epsilon$, across extensive regions, could be attributed to the measurement uncertainty alone

Dvurecenska et al. also state that their metric can only deal with cases where a minimum of six shape descriptors are needed to describe the underlying data field. They suggest that in cases where this is not possible, simpler approaches can be used for the comparison of simulations to measurements. However, they do not explicitly state the approaches one could resort to or provide any guidance on what decision makers should do in those cases.

Finally, none of the aforementioned techniques incorporate the means to tackle

situations where the measurement uncertainty is spatially varying. It will be demonstrated in the next section that the measurement uncertainty can take complex forms; these forms may greatly reduce the space in the feature vector domain of what is considered to be representative of the measurement, information that should be taken into account when feature-based decisions are made.

5.4.4 The effect of heterogeneous measurement uncertainty on the posterior distribution

The cases demonstrated so far in this chapter had a common feature; the field characterizing the measurement uncertainty was considered to be spatially constant. However, this may not be always the case as shown by the works of Wang et al. [156] and Ke et al. [157] for the characterization of the measurement uncertainty in digital image correlation systems. To demonstrate how spatially varying measurement uncertainty can affect the outcome of a validation process, the u_y displacement dataset that was used earlier is employed again across three scenaria. In the first two, the field of measurement uncertainty was synthetically generated while in the third, the field was determined experimentally by Ke et al. [157]. The brown triangles shown in the following figures correspond to the Monte Carlo simulation results of figure 5.14.

The field of measurement uncertainty used in the first scenario can be seen at the bottom left of figure 5.17 and could be considered a simplification of the one determined by Ke et al. shown in figure 5.20. The reasons behind the selection of this field are twofold: its simplicity, as it can be represented using two shape descriptors; the first and the fourth, and the fact that the u_y displacement field is partially described by the same descriptors. The latter could lead to a better understanding of how do the descriptors that characterise both the measurement and the simulation behave with respect to each other.

The resulting posterior distribution can be seen in the same figure. Strong positive correlations emerge between shape descriptors #1 and #4 and #6 and #13 narrowing the domain over which a simulation can be considered representative of the measurement. Moreover, the standard deviations characterizing the magnitude of measurement uncertainty in each shape descriptor have been greatly

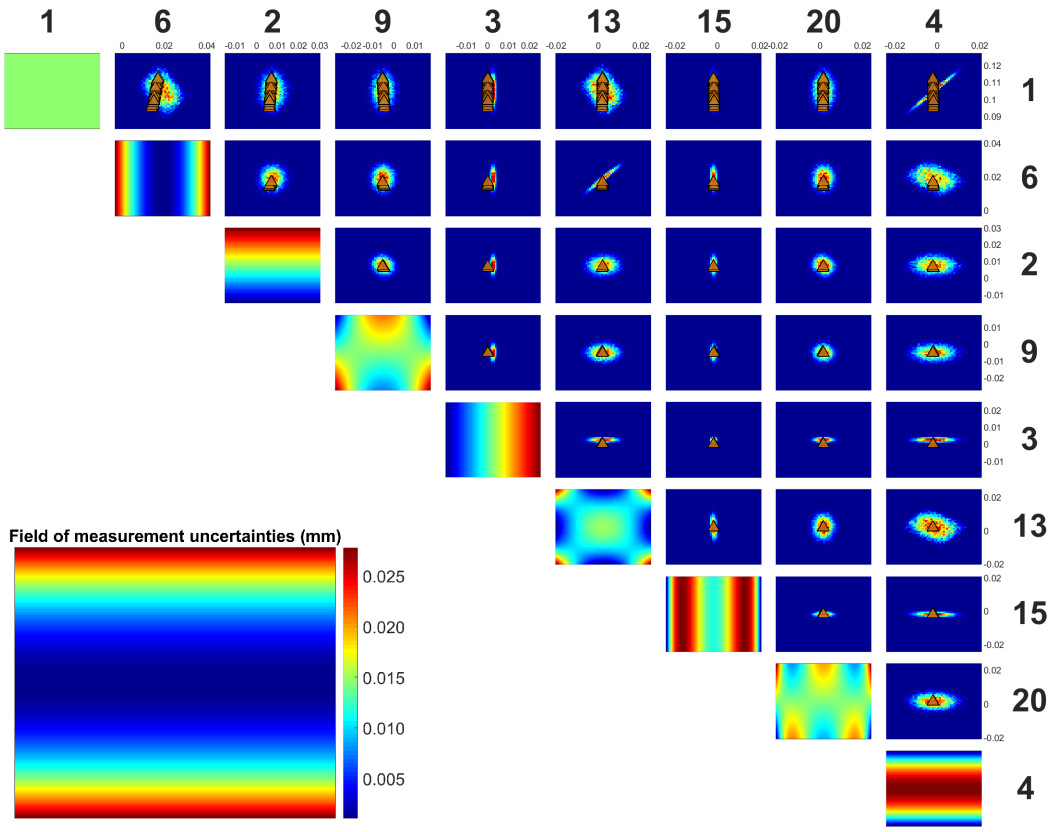

Figure 5.17: Scenario 1: the posterior distribution reflecting the measurement uncertainty in the feature vector space, which is spatially varying, as shown by the inset for the u_y displacement dataset of ROI I. The brown triangles correspond to the Monte Carlo simulation outputs shown in figure 5.15. The change in the form of the posterior distribution is evident even though the spatial average of the measurement uncertainty field remains equal to 0.01 mm.

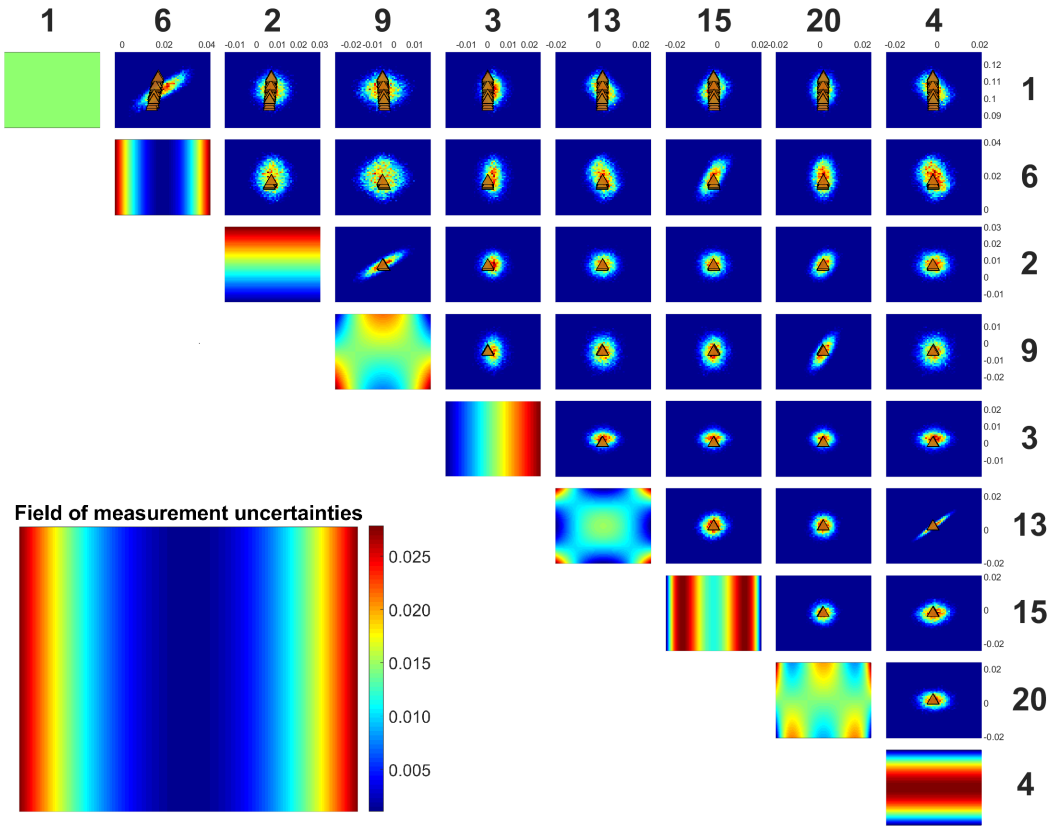

Figure 5.18: Scenario 2: the posterior distribution reflecting the measurement uncertainty in the feature vector space, which is spatially varying, as shown by the inset for the u_y displacement dataset of ROI I. In this case the field of measurement uncertainty is based on shape descriptor #6. It can be seen that shape descriptors #6 and #1 are strongly correlated.

reduced in shape descriptors #3 and #15 as it can be qualitatively observed.

The process described above was repeated in the second scenario with the only difference being that the field of uncertainties was made up of shape descriptors #1 and #6. The result is shown in figure 5.18. Like in the previous case, strong correlations emerge between shape descriptors #1 and #6 both of which are used to describe the displacement and uncertainty fields. One major difference compared to the previous result is that the standard deviations of the posterior marginals have not been reduced as it can be seen from the covariance matrix in figure 5.19.

It is natural that the results of the third scenario depicted in figure 5.20 for the uncertainty field experimentally acquired by Ke et al. [157] will closely resemble the ones of the first scenario given their similarity. The obvious difference in this case shown in figure 5.20 is that the area covered by the posterior distribution is larger, reflecting the larger magnitudes across the uncertainty field. Comparing

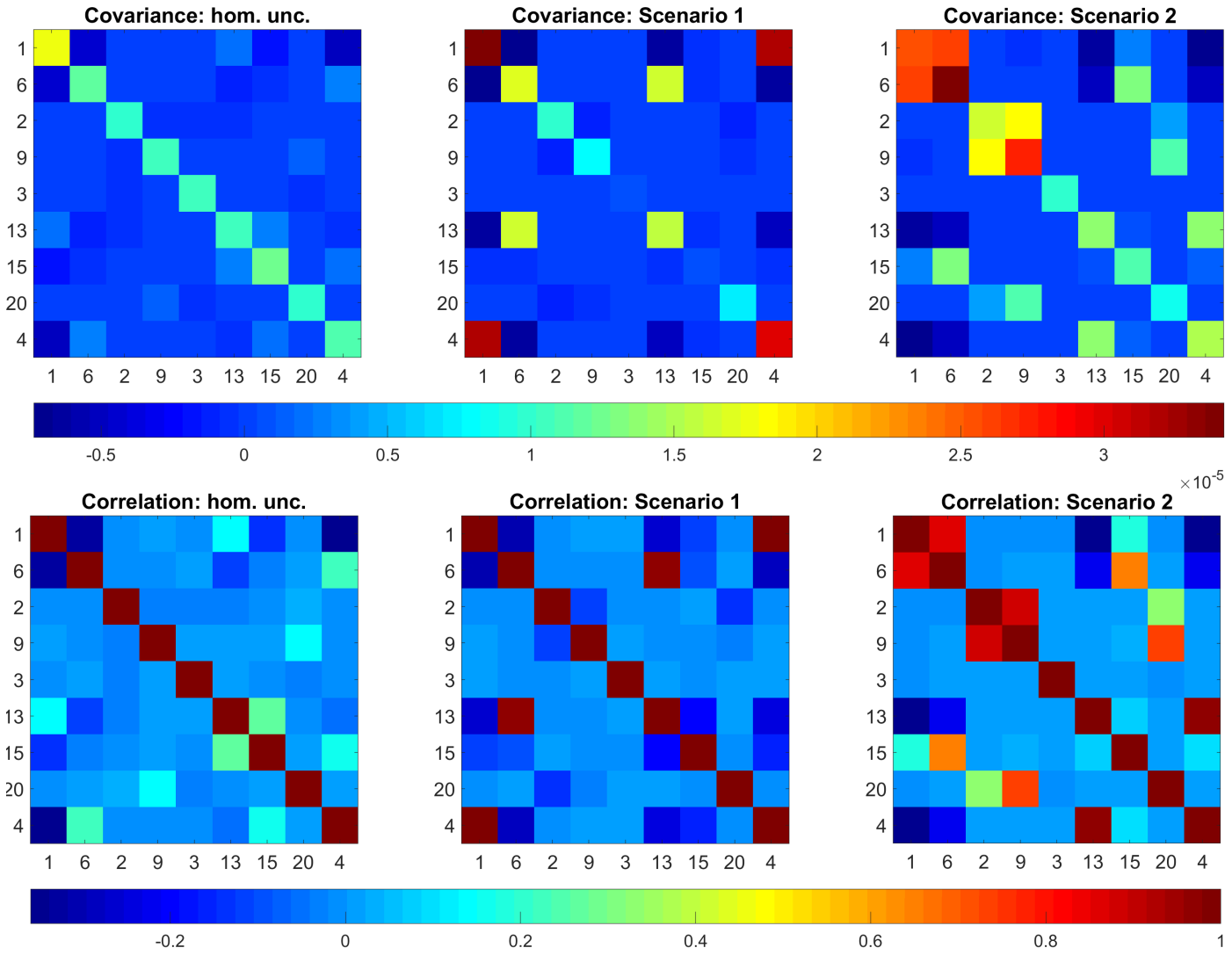

Figure 5.19: Covariance and correlation matrices representing the homogeneous measurement uncertainty for the u_y displacement dataset (ROI I) (left side) and the heterogeneous measurement uncertainty of scenaria 1 (middle) and 2 (right).

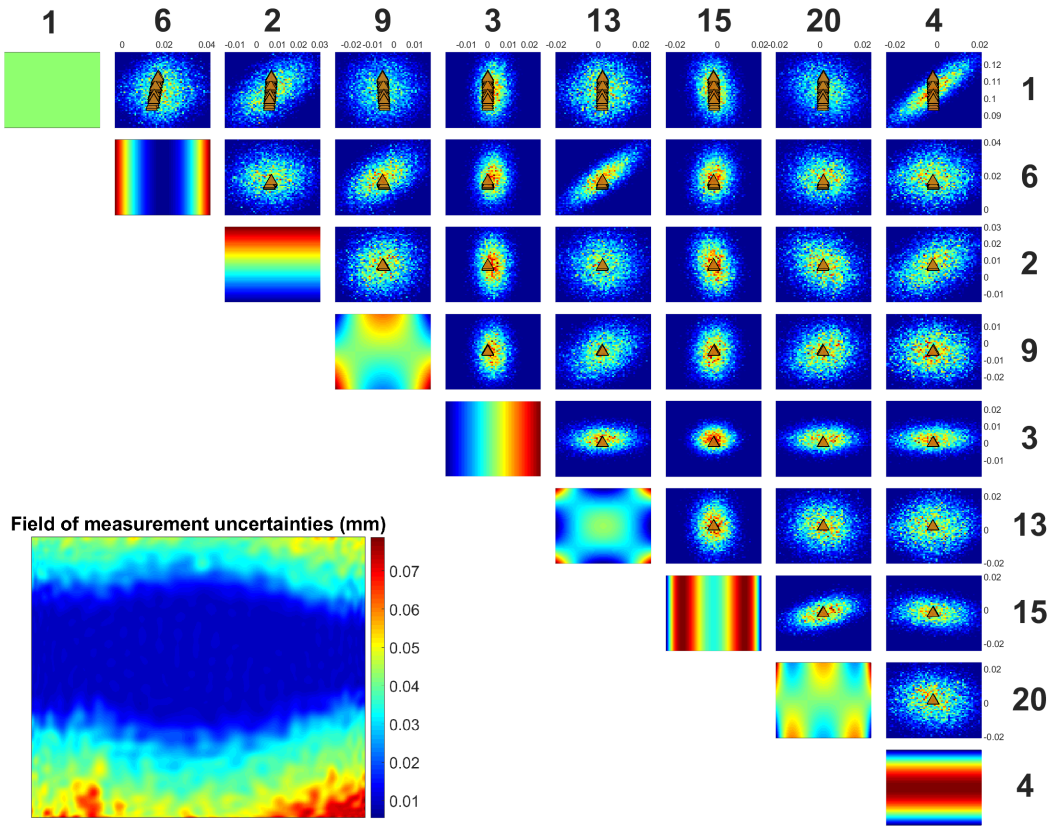

Figure 5.20: Scenario 3: the posterior distribution reflecting the measurement uncertainty in the feature vector space, which is spatially varying, as shown by the inset for the u_y displacement dataset of ROI I. The field of uncertainties was digitized using data by Ke et al.[157] and its form is similar to the one in figure 5.17. As expected, the form of the posterior distribution is qualitatively similar to the one in figure 5.17.

the correlation matrices of the first and third scenario in figure 5.21 it is easy to identify that strong correlations emerge between descriptors #1 and #4 and #6 and #13 in both datasets. What is also interesting in the same figure, is the multitude of emerging correlations such as the ones between shape descriptors #1 and #2 and #6 and #9 that did not exist in the first scenario. This result should be attributed to the non-symmetric, noisy uncertainty field of Ke et al. [157], information which is not included in the synthetically generated field of scenario 1.

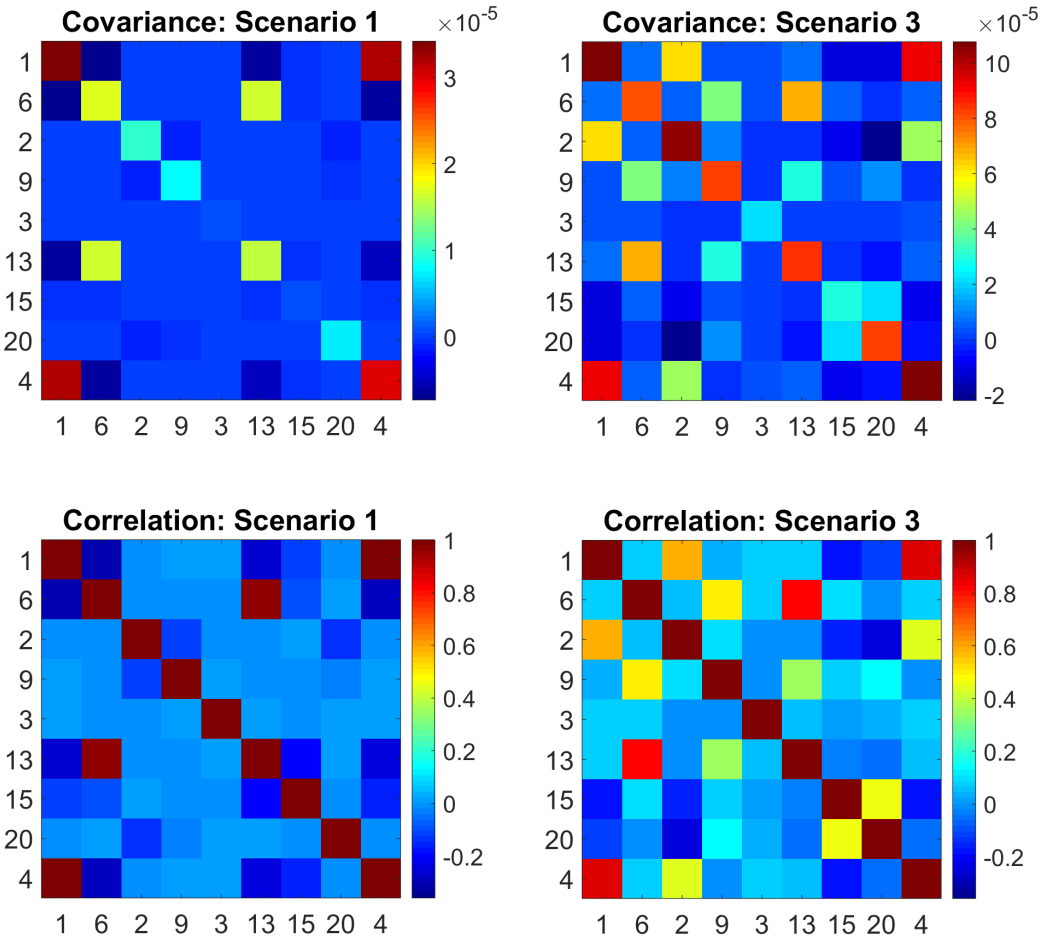

Figure 5.21: Covariance and correlation matrices for the uncertainty fields shown in figures 5.17 (scenario 1) and 5.20 (scenario 3).

5.5 Conclusions

Two novel methods have been developed for the characterization of the validity of a model's predictions which is achieved through their comparison with field measurements infected with uncertainty. They both employ feature extraction techniques and specifically orthogonal decomposition with Chebyshev polynomials as the means to resolving data processing issues. In the first case, the assessment is achieved via a pixel-wise comparison of the predicted field to the experimental measurement. This results in a probabilistic statement for the validity of the former, while accounting for the uncertainty in the latter. In the second case, the validity of a model's prediction(s) is determined through the transformation of the measurement into a probability distribution initially and the calculation of the Mahalanobis distance between that distribution and the prediction's feature vector subsequently.

The conclusions that can be drawn from the use of the proposed methods are the following:

- the probabilistic measure is simple to understand and has the potential to be implemented for decision making. The Mahalanobis distance on the other hand provides an unbounded metric that could be better placed to determine the best among competing models.
- they account for the measurement uncertainty, which may be spatially varying or constant.
- they can be used to provide insight to experimentalists and modelers regarding regions or shape characteristics where the model underperforms.

In addition to the aforementioned, a series of examples using measurements from previously published aerospace engineering tests have been used to demonstrate the applicability of the developed techniques to real-world problems. The measurements, ranging from linearly varying displacements to complex-shaped deformations, act as an indicator of the robustness of the proposed methods.

Chapter 6

Discussion

This chapter will act as the link between the previous chapters where the main body of research was developed and the conclusions where its outcomes will be outlined. The significance of the contributions made in the area of model validation will be discussed and a comparison with the existing techniques will help readers identify their novelties and determine their rigour.

6.1 Validation metrics

Model validation can be considered an emerging and autonomous field aimed at assessing any theory or model used to make predictions. Model assessments have mostly been qualitative, for example through visually comparing predictions against measurements across a series of spatial or temporal locations. The result from such assessments usually takes the form of statements as 'satisfactory', 'good' or 'excellent level of agreement'. Of course, quantitative comparisons are also employed, with the aid of a metric. Even though a significant number of metrics have been proposed, an intuitive understanding of their output along with advantages and disadvantages for some of the most popular ones is still missing.

This gap was partially addressed in the third chapter where the use of some univariate (area metric, u-pooling) and multivariate metrics (the probability integral transform and Mahalanobis distance area metrics) was demonstrated in the context of probabilistic model validation. The presence of parameter uncer-

tainty can influence the outcome of a validation procedure and for that reason it is important to use metrics that can account for that uncertainty. It was shown that the area metric can accurately assess the discrepancy between two univariate probability distributions, demonstrating almost equal sensitivity to differences in the means and the standard deviations of two normal distributions. This capacity combined with the fact that the engineering units are retained during the comparison makes the prospect of using the area metric in probabilistic model validation quite enticing. Bootstrapping was also used to determine 95% confidence intervals for the area metric calculations. This technique provided accurate intervals when the number of samples was large. However, the discrete nature of the empirical distribution function meant that these intervals could be misleading when the number of samples was small, as shown in figure 3.11.

On the other hand, u-pooling, assessed in its capacity to distinguish differences between two distributions is characterised by a more complicated behaviour. This was demonstrated in figure 3.14 where it can be seen that the value of u-pooling does not vary monotonically, relative to the parameters of the underlying distributions, as does the area metric. Given that its value is bounded in the range of $[0, 0.5]$ it makes it hard to intuitively assess what consists of a good or a bad model; for example in the right side of figure 3.37 it can be concluded that two distributions with the same mean but with significantly big differences in their standard deviations (e.g. $\sigma_{exp} = 50, \sigma_{sim} = 0$) result in a u-pooling value of around 0.25. On the other hand the result from the comparison of two distributions where $\Delta\sigma = 25, \mu_{exp} = 25$ and $\mu_{sim} = 10$ is a value less than 0.1 as shown in the bottom left of figure 3.38. The lack of sensitivity demonstrated during the comparison of distributions with profound differences should deter its use.

In addition to these issues, the probability integral transform area metric which comprises the extension of u-pooling in the multivariate space also suffers from dimensionality-related problems. In the case of leftward bias between two distributions, the EDF of the transformed measurements is the indicator function as shown in figure 3.17; this phenomenon, combined with negatively correlated simulations can lead to severely flawed results. Moreover, the capacity of the PIT area metric to efficiently identify differences between distributions is diminished

when the dimensionality of the problem exceeds three as demonstrated in figure 3.18.

Many of these problems are alleviated through the adoption of the Mahalanobis distance area metric. It was demonstrated across different examples that it can accurately quantify differences in the means and standard deviations of multivariate distributions without being influenced by the dimensionality of the problem. It was also shown that it can accurately account for correlations between the different variables without lessening its sensitivity. This feature is significant when assessing models whose response quantities may be correlated like the case of the hole-in-plate experiment. Failing to account for these correlations may result in flawed or inflated results. One limitation associated with its use though, is the assumption that the reference distribution should be normally distributed. This assumption enables the use of the Mahalanobis distance as it allows the reference distribution to be uniquely described by its covariance matrix and its mean. However, when that distribution does not conform to that assumption the outcome could be flawed.

Finally, a demonstration of how these multivariate metrics can be used for probabilistic model validation of full-field measurements via their feature vector form took place. Once again, it became evident that the PIT area metric cannot be confidently used to assess differences between two multivariate distributions. This was demonstrated in the case of u_y displacement data, depicted in figures 3.30 and 3.33, where the combination of dimensionality and negative correlations across shape descriptors resulted in a value of 0. These issues were alleviated by the use of the Mahalanobis distance area metric as shown in figure 3.36. A slight disadvantage though, is the wide range of the output distances as can be seen in the right side of the same figure. Compared to the left figure, the MD-transformed predictions now range in the thousands, a result attributed to the high degree of correlations in the predicted shape descriptors along with a mismatching, high level of variance in the measured shape descriptors. In a manner similar to the 2-dimensional example of figure 3.17 the measurements that do not follow the tightly-defined ellipses in each shape descriptor combination are highly penalised.

Overall, the innovations from this review compared to that of the papers of

Liu et al. [40] and Ling and Mahadevan [158] lie in the following:

- Expanded the analysis of the area metric and u-pooling while providing visualisations of their output across multiple parameter combinations
- Provided an overview of the capacities and limitations for the PIT and MD area metrics
- Suggested a novel method to quantitatively assess the accuracy of stochastic models exploiting full-field measuring capabilities with the aid of decomposition techniques

6.2 Transformation of measurement uncertainty in feature vector space

Even though the aforementioned techniques can be used to assess probabilistic model predictions of spatial data in their decomposed form, these do not account for the effect of measurement uncertainty during that comparison. However, that information should be taken into consideration when decisions are based on their feature vector representation. Although various attempts have been made to represent the uncertainty or variability accompanying a spatial dataset in the feature vector space, an accurate representation is missing. As evidenced by figures 2.4 and 5.16, the attempt made by the CEN committee [8] to account for the effect of measurement uncertainty overrepresented its magnitude, deeming models representative of reality, even when it is apparent from the calculation of their differences that this is erroneous.

This drawback was alleviated with the technique proposed in chapter 4 that makes use of the approximate Bayesian computation. There, the process of iteratively evaluating synthetically generated datasets against the original measurement, in the feature vector space, results in a posterior distribution that represents the measurement and the associated uncertainty. This distribution can be then used to calibrate computational solid mechanics models as in figure 4.4 or to draw inferences regarding the similarity of temporally evolving phenomena such as the El Niño Southern Oscillation depicted in figure 4.14. An advantage stemming

from this technique is that the uncertainty associated with each shape descriptor can be explicitly visualised; jointly through series of scatterplots and histograms shown in figures 4.7 and 5.10 respectively or marginally as in figure 5.16. These visualizations establish a simple tool that can be used to optically determine the proximity between spatial measurements and predictions, and can be considered an improvement to the CEN equivalent.

The suggested technique can also be used to represent spatially varying uncertainty in the feature vector space. Compared to the CEN guide for the case of structural measurements, where the measurement uncertainty was assumed to be spatially constant, this development can be applied to a variety of problems where this many not be the case. This has been demonstrated across a series of examples; from soil moisture measurements to ocean temperature fields and structural measurements. It also becomes apparent from figures 5.14 where the measurement uncertainty is spatially constant and figure 5.17, where the measurement uncertainty is spatially varying, for the structural component case, that the conclusions drawn regarding the validity of a model can be greatly influenced by the form of the measurement uncertainty, even if its spatial average remains constant.

Finally, compared to geostatistical approaches that could be employed to draw and subsequently decompose samples from the underlying stochastic process the proposed technique possesses some advantages:

- Samples are generated iteratively from the underlying distribution without having to undergo the process of detrending and variogram fitting [81] which is the case in traditional geostatistical modeling. Moreover, the assumption of stationarity, which is a requirement in geostatistics is waived, effectively transforming the process into a black box operation.

- The samples that are generated from the posterior distribution during the ABC, reflect the measurement uncertainty only for the selected components or shape descriptors that were initially selected to characterise the measurement. This is computationally more efficient compared to the geostatistical approach, where drawing samples from the underlying stochastic process

and then decomposing them would mean the inclusion of shape descriptors other than the ones already selected.

6.3 Model validation using full-field measurements

Chapter 5 can be considered the pinnacle of this research. Developing a method to accurately assess the quality of a model's predictions and demonstrating that on a test case while exploiting spatial measurements has been the aim of the research. This could have been achieved earlier with the use of the Mahalanobis distance area metric. However, the presence of significant amounts of measurement uncertainty, combined with the lack of a method to accurately transform that uncertainty in the feature vector space presented the need for the development of one. Having tackled that challenge, the question that remained was how to quantitatively assess the quality of a prediction while accounting for the measurement uncertainty. A twofold answer to that question has been given:

- A probabilistic metric described by equation (5.1), similar to the reliability metric [76] where the quality of the model's prediction is assessed via a series of pixel-wise comparisons with the measurement, while accounting for the measurement uncertainty.
- A metric based on the Mahalanobis distance between a measurement and prediction(s), described by equations (5.5) and (5.6) when the validation is performed in the feature vector space. This forms a way to assess the quality of the model while accounting for the measurement uncertainty in the feature vector space.

Both techniques were applied to a series of structural mechanics examples and their advantages and novelties will be discussed. A significant improvement, with respect to structural model validation compared to the CEN guide, is that instead of an accept/reject statement, the proposed techniques have the capacity to provide quantitative assessments of the model's quality when compared to a measurement. These assessments take two forms: a percentage describing

the proportion of pixel-wise differences that can be attributed to the existence of measurement uncertainty, and an unbounded metric quantifying the distance of a prediction from a multivariate distribution, in this case corresponding to the feature vector form of the measurement and its associated uncertainty. The benefit of the former metric lies in its simplicity, allowing vast amounts of information to be compressed and communicated in a simple intuitive manner to non-experts, while that of the Mahalanobis distance is that it accounts for the measurement uncertainty in the feature vector space thus allowing decisions to be drawn in the same space.

Compared to the probabilistic metric whose output can be quite easily interpreted as a simple frequency, the outcome of the Mahalanobis distance is not that intuitive; especially when this is used to calculate distances in high-dimensional spaces while being normalised by the covariance matrix. The question that naturally arises is what constitutes a prediction representative of the measurement and its uncertainty?

To answer that question the idea implemented in the MD area metric of Zhao et al. [66] was adopted. In this case, the MD of every sample drawn from the posterior distribution is calculated with respect to the mean of the posterior itself. This results in a distribution of distances, ranging from zero when a sample is located on the mean of the distribution, extending to an upper value that corresponds to the most distant sample. To capture that information, the histogram of the calculated Mahalanobis distances is initially plotted; afterwards, the Mahalanobis distances of the simulations are calculated with respect to mean of the posterior and plotted on the same figure. Comparing the Mahalanobis distances of the simulation(s) with respect to the range corresponding to the measurement itself should nominally provide an indication of whether the prediction could be considered representative of the measurement or not. However, this conclusion is correct under the assumption that the posterior distribution is normally distributed, which may not be always the case as shown in figure 5.13. Examples of this caveat and how to handle it are shown in figures 5.13 and 5.15 respectively.

Worden and colleagues [105], [107], [155] developed a Mahalanobis-based technique much earlier to evaluate the quality of a model in the presence of uncertain-

ties. However, their technique has a significant difference compared to the one proposed here. The characterisation of the measurement and its uncertainty in the feature vector domain has been performed using ABC. This procedure limits the assumptions about the form of the measurement error (such as normality) enabling the representation of an information-rich spatial field in the feature vector domain.

Even though the output of the validation metric is primarily used to inform decisions during model validation, the 'by-products' of the process can be used to extract more information about the problem at hand; pixel-wise differences can be used to identify where the largest model-experiment deviations occur in the region of interest, thus allowing modelers to identify regions where their models fail to represent the real world. In the case of feature-based model evaluations, a visual comparison of the measurement and its uncertainty against the predictions as in figures 5.10 and 5.11 can be used to identify shape descriptor combinations where the largest deviations occur and potentially lead to a better understanding of the underlying physics.

Various techniques have been proposed to tackle model validation using spatial measurements in engineering. Dvurecenska et al. [11] developed a relative error metric that can be used to assess the similarity of two datasets in their feature vector form while accounting for the measurement uncertainty via a normalised error threshold. A detailed explanation of the procedure along with an illustration are given in the literature review and in figure 2.8 respectively. A comparison of the output of the proposed metrics with the relative error metric for the case of the I-beam data is given in table 5.3. The conclusions that can be drawn from the comparison are the following: i) the misrepresentation of the measurement uncertainty by the relative error metric. This stems from the fact that the metric results in an impressive value of 100% for the case of ϵ_χ in ROI 2, whereas witnessed by figures 5.4 and 5.7 that value seems highly unlikely to be realistic. This statement is also supported by the synthetic example of figure 2.4 where the outcome of the relative error metric is also 100% even though it is obvious that there are extensive regions where the local differences exceed more than five times the measurement uncertainty; ii) the limitation to cases where the measurement

uncertainty is spatially constant as reflected by equation (2.15); iii) the inability to characterise dissimilarities for cases where the number of shape descriptors used to describe the underlying dataset is less than 6 as in the case of ROI 2, u_y displacements. Even though the authors suggest that for data fields in which the variable is a nonlinear function of both spatial coordinates at least 6 shape descriptors are needed to accurately describe it, this point may not alway be valid as was shown by Berke et al. [123]. In that case modal analyses of Hastelloy-X plates resulted in non-linear mode shapes for which a small number of shape descriptors was used to accurately describe them.

Lampeas et al. [9] proposed the use of Lin's concordance correlation coefficient (CCC) [159] between the measured and the simulated shape descriptors to quantify the quality of the prediction in addition to the accept/reject outcome they acquired from the application of the CEN instructions. The CCC can be considered an improvement of Pearson's correlation coefficient in that can account for scatter, scale and location shift between two sets of observations. Similar to Pearson's correlation coefficient it is bounded in the range of $[-1, 1]$, with the latter suggesting perfect similarity. Even though the use of the concordance correlation coefficient leads to a quantitative outcome, the decision based on that outcome can be severely flawed. In the case of the datasets depicted in figure 5.2 for the I-beam, the authors found that the concordance correlation was greater than 0.98 (suggesting perfect agreement) for all the datasets except the ROI 1, ϵ_χ dataset whose value was about 0.87. This can be explained in a manner similar to the one for the datasets depicted in figure 2.4 where the correlation coefficient between the two feature vectors is $\rho = 0.965$ and appears contradictory to the corresponding pixel-wise differences where much larger deviations are demonstrated. This phenomenon stems from the fact that local pixel-related information is lost when the comparison is based on the shape descriptor values and not on the raw measurements. Similarly, Allemang [10] used the slope between the predicted and measured singular values to characterise the model's quality (at least for the case of the first principal component). To demonstrate why the use of correlation metrics may not be the best approach to assess the similarity between two spatial datasets based on their feature vector form an example is given; a total of 30.000

simulations, similar to the one of figure 2.4 were performed. The u_y dataset of figure 2.2 was once again used as the measurement, while the simulations consisted of random perturbations of the measurement's feature vector. The outcome from the assessment of each simulation against the measurement can be seen in figure 6.1. On the top left, the correlation coefficient (Pearson's) between the reference and the perturbed shape descriptors has been plotted in the x-axis against the mean absolute error resulting from the calculation of their pixel-wise differences. The latter is defined as:

$$MAE = \frac{\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} |Meas(i,j) - Sim(i,j)|}{N_x N_y} \tag{6.1}$$

where N_x and N_y correspond to the total number of pixels in the horizontal and vertical direction, $Meas(i,j)$ is the measured dataset, $Sim(i,j)$ is the perturbed dataset and i,j are the pixel locations. The MAE was selected as a simple heuristic to aid the user in conceptualizing the magnitude of pixel-wise differences given that $u_{meas}(i,j) = 0.01$ mm.

It can be seen that a correlation coefficient of $\rho = 0.95$ between the two feature vectors is attained when the mean absolute error ranges between 0.02mm and 0.05mm with the latter reflecting an average difference of 0.05 mm between the predicted and measured datasets. This means that their difference is on average five times the measurement uncertainty ($u_{meas} = 0.01$ mm), implying that it could locally be much larger and could not be attributed to the presence of measurement uncertainty alone. Similar statements can be made for larger values of ρ indicating that conclusions of the type: 'a correlation coefficient of 0.99 suggests perfect agreement between the two datasets', should be made with caution when they are based on a transformed form of the initial data.

The amount of scatter depicted on the top left side of the figure is evidently reduced on the right side, where the similarity between the two feature vectors has been calculated using Lin's concordance correlation coefficient. In this case a CCC of 0.95 corresponds to mean differences ranging between 0.02 mm and 0.03 mm, result that poses a great reduction in vertical scatter compared to Pearson's correlation. Even though the results stemming from the use of the CCC in this example demonstrate less scatter to the ones of Pearson's correlation, the

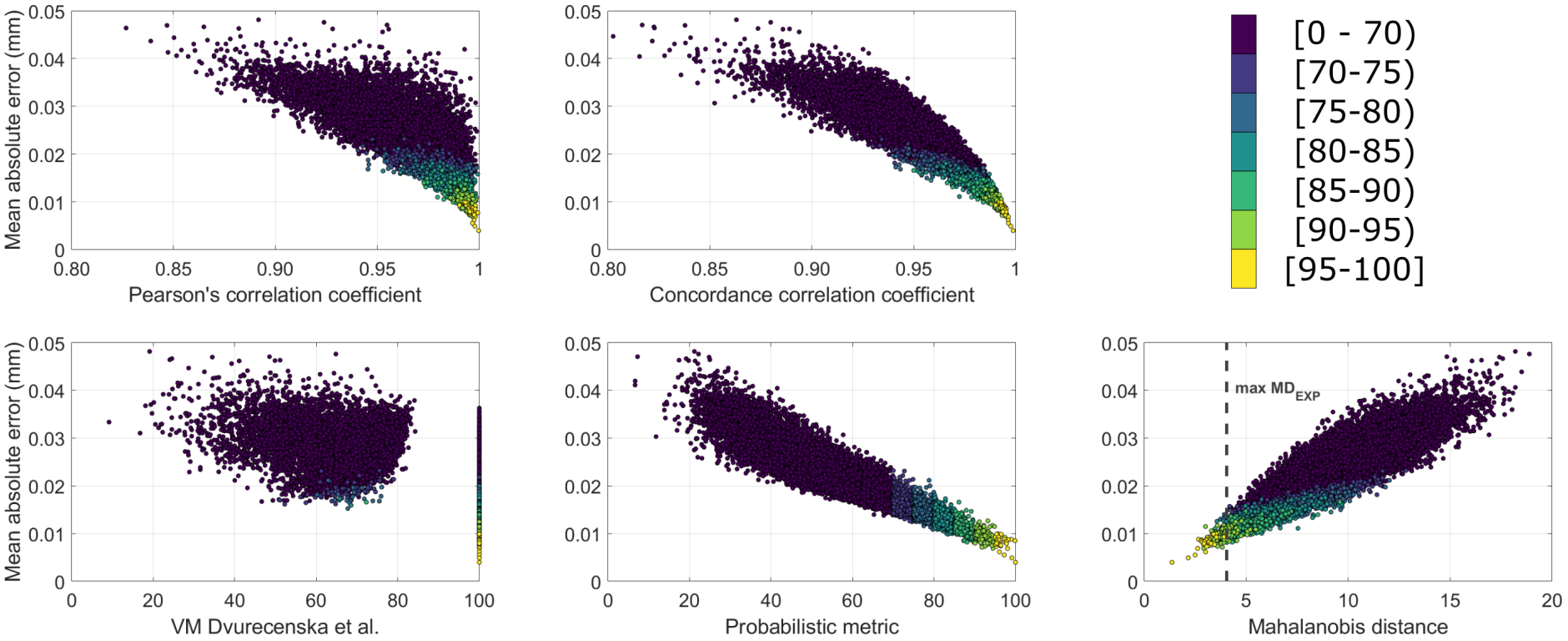

Figure 6.1: A variety of metrics used to assess the similarity across spatial fields is depicted. A total of 30.000 Monte Carlo perturbations of the u_y displacement dataset shown in figure 2.2 are assessed against the initial dataset. The selected colour bands correspond to the results of the probabilistic metric. Given that the measurement uncertainty is equal to 0.01 mm the yellow-coloured circles (95%-100%) correspond to simulations that the probabilistic metric considers representative of the measurement. The results from the feature-based assessments (all except the probabilistic metric which is pixel-based) and the probabilistic metric have been plotted against the mean absolute error (pixel-based). For the case of the Mahalanobis distance, the vertical dashed line corresponds to the maximum distance that could deem a prediction to be representative of the measurement.

conclusions that can be drawn from its use in model validation are somewhat limited. Analytically: i) it is impossible to determine what constitutes a good model using the CCC alone; ii) it is not obvious how the measurement uncertainty can be accounted when calculating the CCC.

In addition to the correlation coefficients, the outcomes from the assessment of the simulated datasets against the measurement using the validation metric proposed by Dvurecenska et al. and the two novel metrics have been plotted in the same figure. The colouring used across the graphs depicts the assessment using the probabilistic metric at each point. Compared to the majority of the measures depicted in this figure, it seems that the metric proposed by Dvurecenska et al. demonstrates a rather unconventional behaviour. It is apparent that a large portion (almost two thirds) of the 30.000 simulations are considered to be 100% representative of the measurement forming a big cluster of simulations rated 100% with the next proximate value to be around 80%. On the other hand, the probabilistic metric follows a pattern of increasing values as the mean absolute error is reduced. This results in a total of 31 simulations to be considered representative of the measurement. Finally, on the right side, the Mahalanobis distance of the simulations with respect to the measurement is shown along with the upper threshold depicted by the grey, dashed, vertical line. A total of 152 simulations are deemed representative of the measurement. Compared to the 31 simulations reported earlier, this represents a significant increase and can be attributed to the deviation from normality of the distribution corresponding to the measurement.

From this example it can be concluded that: a) feature-based assessments using commonly employed correlation metrics such as Pearson's or Lin's correlation coefficients can be misleading b) the Mahalanobis distance-based metric could lead to false inferences regarding the capacity of a model to represent the real world if the distribution corresponding to the measurement and its uncertainty is not Gaussian.

Chapter 7

Conclusions and suggestions for future research

7.1 Conclusions

As computational power rapidly increases so does the need for models capable of simulating the real world. The outcomes of these models are used to inform decisions with potentially substantial socio-economic impact and their capacity to represent the real world should be ascertained. To determine the level to which a model represents the real world, some form of model assessment is required. This process, which is known as model validation has been the focus of research and the resulting developments can be used to assist decision makers faced with this information abundance. Even though this research has been largely focused on the validation of computational solid mechanics models, it becomes apparent from the range of the examples and applications that the novelties of the proposed techniques can be easily transferred to any scientific discipline where spatial, gridded measurements are available. The major contributions of this research are the following:

An extensive review of some popular validation metrics used to assess probabilistic models. The review included both univariate and multivariate metrics. The novelty of this exercise lies in the fact that for a first time an analysis of the behavior of those metrics across a wide range of con-

ditions took place. It was concluded that the area metric is equally sensitive to differences in the means and the standard deviations of two normal distributions representing point measurements and simulations. U-pooling on the other hand is characterised by a more complicated behavior and a lack of sensitivity in identifying differences between distributions. Similarly, the probability integral transform area metric, which extends the use of u-pooling in the multivariate space demonstrates limited capability in correctly assessing differences between distributions when the dimensionality of the problem exceeds three, or when negative correlations among the variables emerge. These issues are alleviated by the Mahalanobis distance area metric which can accurately identify dissimilarities between measurements and predictions even in high dimensions.

The development of a a method to validate stochastic models when multiple spatial measurements are available. This is based on feature extraction techniques, to transform data-rich spatial predictions and measurements into points belonging to a lower-dimensionality space. The collection of simulation outputs is then compared to measurements using the Mahalanobis distance area metric. This comparison produces an output that even though quantitative, would be better suited in informing the selection of the best model among competing ones. Moreover, the measurement uncertainty upon which an extensive amount of work has been focused on, is not taken into consideration during this comparison. Despite its limitations this method constitutes an innovation in the area of model validation where a way to jointly assess multiple spatial simulation outputs against measurements is currently lacking.

The development of a method to transform the measurement uncertainty characterising a spatial measurement in its feature vector form. This is achieved via a Monte Carlo approach where the spatial measurement is iteratively evaluated against synthetically generated datasets. The feature vectors corresponding to the synthetic datasets that were previously deemed representative of the measurement comprise samples that make up a multivariate distribution. This distribution represents the spatial measurement and its uncertainty in the feature vector space. This alternative representation of the measurement and its uncertainty allows for statistical inference methods to be applied, solely on the

feature vector space, a characteristic that is particularly attractive for temporally consecutive measurements where the identification of critical events is the aim.

A benefit of this development is its capability to be used with different feature extraction techniques and spatial measurements of varying complexity as demonstrated by the numerous applications. These ranged from linearly varying fields of displacements that were decomposed using orthogonal, Chebyshev polynomials to oceanographic datasets which are described by non-linear dynamics and principal component analysis was employed. The provision of an explicit definition of the extent of the uncertainty in a feature-transformed space is an improvement to the currently established practices, where such a definition is missing.

Two novel approaches have been proposed to quantitatively assess a model's quality while accounting for the uncertainty in the measurements. The first via the calculation of the percentage of pixel-wise differences that are within the measurement and its uncertainty. This is realized with the aid of orthogonal decomposition techniques to homogenize measurements and predictions that lie in different grids. This metric, which is bounded between 0% and 100% provides an intuitive and simple way to convey model assessments, based on data-rich spatial information, to decision makers. **The second technique is based on the calculation of the Mahalanobis distance between a feature vector and the mean of a distribution; the former corresponding to a simulation or an ensemble of simulations and the latter representing a measurement and its uncertainty.** In this case the assessment of the model's quality is solely determined on its feature vector form, resulting from some decomposition technique, while the information concerning the measurement and its uncertainty is reflected by the mean and the covariance matrix of the distribution.

Both techniques can be used along with a field of constant or spatially varying measurement uncertainties, while providing a simple way to convey information regarding the quality of the prediction(s) to decision makers. Various visualizations have been employed to aid that process and better inform modelers and experimentalists of regions or features where the model underperforms.

The proposed techniques also benefit from the fact that only a small num-

ber of assumptions are made regarding the structure of the data, thus allowing them to be used in a black-box setting. This poses a considerable simplification compared to traditional geostatistical techniques where numerous and often unfounded assumptions must be met to allow further analysis.

7.2 Future work

It has been repeatedly stated that model validation is the process associated with determining the degree to which a model represents the real world. This process which is largely dependent on the use of a validation metric to assess the level of (dis)agreement between measurements and predictions cannot be considered complete until a decision to accept or to reject a model's predictions is made. This decision is based on some pre-defined accuracy requirement, which is usually a function of the significance of the comparison and of the impact of making mistakes. The accuracy requirement implemented throughout this thesis has been based on the measurement uncertainty accompanying a field of measurements in a manner similar to that of the CEN guide [8] allowing the characterisation of a distribution in the feature vector space that describes the measurement and its uncertainty. However, **an explicit characterisation of what constitutes a simulation acceptable**, in the form of a percentage (for example 5%-10% of the measured quantities), **would allow decision makers to separate the effect of measurement uncertainty from the accuracy requirements**. This could be achieved by modifying the existing ABC algorithm while supplementary visualizations would provide a more intuitive understanding of the effect of uncertainty and accuracy requirements and would better inform the final decision.

Although the developed techniques make use of the entire field of measurements and predictions available, they do not **account for the amount of information redundancy in the data**. This redundancy which stems from spatial autocorrelation, meaning that neighbouring values are similar, violates the traditional assumption of independence (in statistics) across measurements and must be taken into account when the aim is to determine the level of association

between two datasets (simulation and measurement) with a certain degree of confidence. Achieving that **would maximize the potential stemming from modern full-field measuring devices to making data-driven decisions**. This could build on the works by Clifford et al. [83] and Griffith [79] on the effective degrees of freedom for spatial data to characterise the confidence about the significance of correlations between measurements and predictions or about equality between two response quantities (such as the means between the predicted and measured fields).

A glimpse of how the transformed form of the measurement and its uncertainty in the feature vector space, could be used for model calibration was given in figure 4.4. It was shown that data-rich spatial measurements can be combined with model predictions to determine the values of material constants that minimize the measurement-prediction error. It would be of interest to **demonstrate how this technique could be used to extract material information**, especially in the case of composite materials where a multitude of tests is needed to fully characterize them. An overview of existing methods exploiting full-field measurements for mechanical parameter identification can be found in [160]. In addition to these methods, Wang et al. [103] and Gogu et al. [161] have developed feature-based parameter identification techniques and a comparison would be meaningful.

References

[1] Patterson, E. (2015). 'On the credibility of engineering models and meta-models'. *Journal of Strain Analysis for Engineering Design*, 50(4), pp. 218–220.

[2] Schruben, L.W. (1980). 'Establishing the credibility of simulations'. *Simulation*, 34(3), pp. 101–105.

[3] Oberkampf, W.L. and Roy, C.J. (2010). *Verification and Validation in Scientific Computing*. Cambridge University Press.

[4] ASME (2006). *V&V 10-2006, Guide for Verification and Validation in Computational Solid Mechanics*. New York: American Society of Mechanical Engineers.

[5] Roache, P.J. (1998). 'Verification of codes and calculations'. *AIAA journal*, 36(5), pp. 696–702.

[6] Mottershead, J.E. and Friswell, M.I. (1993). 'Model updating in structural dynamics: A survey'. *Journal of Sound and Vibration*, 167(2), pp. 347–375.

[7] Sebastian, C., Hack, E. and Patterson, E. (2012). 'An approach to the validation of computational solid mechanics models for strain analysis'. *The Journal of Strain Analysis for Engineering Design*, 48(1), pp. 36–47.

[8] CEN (2014). *CEN Workshop Agreement CWA 16799:2014, Validation of computational solid mechanics models* . European Committee for Standardization.

[9] Lampeas, G., Pasialis, V., Lin, X. and Patterson, E.A. (2015). 'On the validation of solid mechanics models using optical measurements and data decomposition'. *Simulation Modelling Practice and Theory*, 52, pp. 92–107.

[10] Allemang, R., Kolluri, M.M., Spottswood, M. and Eason, T. (2016). 'Decomposition-based calibration/ validation metrics for use with full-field measurement situations'. *Journal of Strain Analysis for Engineering Design*, 51(1), pp. 14–31.

[11] Dvurecenska, K., Graham, S., Patelli, E. and Patterson, E.A. (2018). 'A probabilistic metric for the validation of computational models'. *Royal Society Open Science*, 5(11).

[12] Kleindorfer, G.B., O'Neill, L. and Ganeshan, R. (1998). 'Validation in simulation: Various positions in the philosophy of science'. *Management Science*, 44(8), pp. 1087–1099.

[13] Popper, K.R. (2005). *The logic of scientific discovery*. Routledge.

[14] Robinson, A.P., Duursma, R.A. and Marshall, J.D. (2005). 'A regression-based equivalence test for model validation: Shifting the burden of proof'. *Tree Physiology*, 25(7), pp. 903–913.

[15] Howson, C. and Urbach, P. (2006). *Scientific Reasoning: The Bayesian Approach*. Open Court Publishing.

[16] Rudner, R. (1953). 'The Scientist Qua Scientist Makes Value Judgments'. *Philosophy of Science*, 20(1), pp. 1–6.

[17] Taleb, N. (2005). *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets*. Random House.

[18] Audi, R. (2010). *Epistemology: A Contemporary Introduction to the Theory of Knowledge*. Routledge.

[19] AIAA (1998). *Guide for the Verification and Validation of Computational Fluid Dynamics Simulations*. American Institute of Aeronautics & Astronautics.

[20] DoD (1996). *Instruction 5000.61, DoD Modeling and Simulation (M&S) Verification, Validation and Accreditation,*. United States Department of Defense.

[21] Beer, M., Ferson, S. and Kreinovich, V. (2013). 'Imprecise probabilities in engineering analyses'. *Mechanical Systems and Signal Processing*, 37(1-2), pp. 4–29.

[22] Worden, K., Barthorpe, R., Cross, E., Dervilis, N., Holmes, G., Manson, G. and Rogers, T. (2018). 'On evolutionary system identification with applications to nonlinear benchmarks'. *Mechanical Systems and Signal Processing*, 112, pp. 194–232.

[23] Alefeld, G. and Herzberger, J. (2012). *Introduction to interval computation*. Academic Press.

[24] Ferson, S. (2001). 'Probability bounds analysis solves the problem of incomplete specification in probabilistic risk and safety assessments'. In 'Risk-Based Decisionmaking in Water Resources IX', pp. 173–188.

[25] Ferson, S., Kreinovich, V., Grinzburg, L., Myers, D. and Sentz, K. (2015). *Constructing probability boxes and Dempster-Shafer structures*. Albuquerque, NM (United States): Sandia National Lab.

[26] Tonon, F. (2004). 'Using random set theory to propagate epistemic uncertainty through a mechanical system'. *Reliability Engineering & System Safety*, 85(1-3), pp. 169–181.

[27] Ferson, S. and Ginzburg, L.R. (1996). 'Different methods are needed to propagate ignorance and variability'. *Reliability Engineering and System Safety*, 54(2-3), pp. 133–144.

[28] Der Kiureghian, A. and Ditlevsen, O. (2009). 'Aleatory or epistemic? does it matter?' *Structural safety*, 31(2), pp. 105–112.

[29] Patelli, E., Govers, Y., Broggi, M., Gomes, H.M., Link, M. and Mottershead, J.E. (2017). 'Sensitivity or bayesian model updating: a comparison of techniques using the dlr airmod test data'. *Archive of Applied Mechanics*, 87(5), pp. 905–925.

[30] Kennedy, M.C. and O'Hagan, A. (2001). 'Bayesian Calibration of Computer Models'. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3), pp. 425–464.

[31] Sisson, S.A., Fan, Y. and Beaumont, M. (2018). *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC.

[32] Bayarri, M.J. and Berger, J.O. (2004). 'The interplay of Bayesian and frequentist analysis'. *Statistical Science*, pp. 58–80.

[33] Smith, R.C. (2013). *Uncertainty quantification: Theory, Implementation, and Applications*. SIAM.

[34] ISO (2017). *ISO/IEC 17025: 2017: General Requirements for the Competence of Testing and Calibration Laboratories*. International Organization for Standardization.

[35] Hack, E., Lin, X., Patterson, E.A. and Sebastian, C.M. (2015). 'A reference material for establishing uncertainties in full-field displacement measurements'. *Measurement Science and Technology*, 26(7), pp. 075004.

[36] BIPM (2008). *CGM 100:2008, GUM 1995 with minor corrections. Evaluation of measurement data—guide to the expression of uncertainty in measurement*. Joint Committee for Guides in Metrology.

[37] Taylor, J.R. (1982). *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. University Science Books.

[38] Oberkampf, W.L. and Trucano, T.G. (2002). 'Verification and validation in computational fluid dynamics'. *Progress in Aerospace Sciences*, 38(3), pp. 209–272.

[39] Oberkampf, W.L. and Ferson, S. (2007). *Model Validation Under Both Aleatory and Epistemic Uncertainty. SAND2007-7163C.* Albuquerque, NM (United States): Sandia National Lab.

[40] Liu, Y., Chen, W., Arendt, P. and Huang, H.Z. (2011). 'Toward a Better Understanding of Model Validation Metrics'. *Journal of Mechanical Design*, 133(7), pp. 071005.

[41] Montgomery, D.C. and Runger, G.C. (2010). *Applied Statistics and Probability for Engineers.* John Wiley & Sons.

[42] Kolmogoroff, A. (1941). 'Confidence limits for an unknown distribution function'. *Annals of Mathematical Statistics*, 12(4), pp. 461–463.

[43] Smirnov, N.V. (1939). 'On the estimation of the discrepancy between empirical curves of distribution for two independent samples'. *Bull. Math. Univ. Moscou*, 2(2), pp. 3–14.

[44] Razali, N.M. and Wah, Y.B. (2011). 'Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests'. *Journal of Statistical Modeling and Analytics*, 2(1), pp. 21–33.

[45] Rebba, R. and Mahadevan, S. (2006). 'Validation of models with multivariate output'. *Reliability Engineering & System Safety*, 91(8), pp. 861–871.

[46] Kruschke, J. (2014). *Doing Bayesian Data Analysis: A tutorial with R, JAGS, and Stan.* Academic Press.

[47] Rebba, R., Mahadevan, S. and Huang, S. (2006). 'Validation and error estimation of computational models'. *Reliability Engineering & System Safety*, 91(10-11), pp. 1390–1397.

[48] Rebba, R. and Mahadevan, S. (2008). 'Computational methods for model reliability assessment'. *Reliability Engineering & System Safety*, 93(8), pp. 1197–1207.

[49] Jeffreys, H. (1998). *The Theory of Probability.* Oxford University Press.

[50] Kass, R.E. and Raftery, A.E. (1995). 'Bayes factors'. *Journal of the American Statistical Association*, 90(430), pp. 773–795.

[51] Oberkampf, W.L. and Barone, M.F. (2006). 'Measures of agreement between computation and experiment: Validation metrics'. *Journal of Computational Physics*, 217(1), pp. 5–36.

[52] Ferson, S., Oberkampf, W.L. and Ginzburg, L. (2008). 'Model validation and predictive capability for the thermal challenge problem'. *Computer Methods in Applied Mechanics and Engineering*, 197(29-32), pp. 2408–2430.

[53] Zhan, Z., Fu, Y. and Yang, R.J. (2013). 'On stochastic model interpolation and extrapolation methods for vehicle design'. *SAE International Journal of Materials and Manufacturing*, 6(3), pp. 517–531.

[54] Bredbenner, T.L., Eliason, T.D., Francis, W.L., McFarland, J.M., Merkle, A.C. and Nicolella, D.P. (2014). 'Development and validation of a statistical shape modeling-based finite element model of the cervical spine under low-level multiple direction loading conditions'. *Frontiers in Bioengineering and Biotechnology*, 2, pp. 58.

[55] Rosenblatt, M. (1952). 'Remarks on a multivariate transformation'. *Annals of Mathematical Statistics*, 23(3), pp. 470–472.

[56] Jung, B.C., Park, J., Oh, H., Kim, J. and Youn, B.D. (2015). 'A framework of model validation and virtual product qualification with limited experimental data based on statistical inference'. *Structural and Multidisciplinary Optimization*, 51(3), pp. 573–583.

[57] Gorguluarslan, R.M., Choi, S.K. and Saldana, C.J. (2017). 'Uncertainty quantification and validation of 3D lattice scaffolds for computer-aided biomedical applications'. *Journal of the Mechanical Behavior of Biomedical Materials*, 71, pp. 428–440.

[58] Gardner, P., Lord, C. and Barthorpe, R.J. (2018). 'An evaluation of validation metrics for probabilistic model outputs'. In 'ASME 2018 Verification

and Validation Symposium', American Society of Mechanical Engineers Digital Collection.

[59] Silva, A.S., Sebastian, C., Lambros, J. and Patterson, E. (2019). 'High temperature modal analysis of a non-uniformly heated rectangular plate: Experiments and simulations'. *Journal of Sound and Vibration*, 443, pp. 397–410.

[60] Li, W., Chen, W., Jiang, Z., Lu, Z. and Liu, Y. (2014). 'New validation metrics for models with multiple correlated responses'. *Reliability Engineering and System Safety*, 127, pp. 1–11.

[61] Genest, C. and Rivest, L.P. (2001). 'On the multivariate probability integral transformation'. *Statistics and Probability Letters*, 53(4), pp. 391–399.

[62] Mahalanobis, P.C. (1936). 'On the generalized distance in statistics'. National Institute of Science of India.

[63] De Maesschalck, R., Jouan-Rimbaud, D. and Massart, D.L. (2000). 'The Mahalanobis distance'. *Chemometrics and Intelligent Laboratory Systems*, 50(1), pp. 1–18.

[64] Brereton, R.G. and Lloyd, G.R. (2016). 'Re-evaluating the role of the Mahalanobis distance measure'. *Journal of Chemometrics*, 30(4), pp. 134–143.

[65] Bi, S., Prabhu, S., Cogan, S. and Atamturktur, S. (2017). 'Uncertainty quantification metrics with varying statistical information in model calibration and validation'. *AIAA Journal*, 55(10), pp. 3570–3583.

[66] Zhao, L., Lu, Z., Yun, W. and Wang, W. (2017). 'Validation metric based on Mahalanobis distance for models with multiple correlated responses'. *Reliability Engineering and System Safety*, 159, pp. 80–89.

[67] Hu, J., Jiang, P., Zhou, Q., McKeand, A. and Choi, S.K. (2020). 'Model validation methods for multiple correlated responses via covariance-overlap based distance'. *Journal of Mechanical Design*, 142(4).

[68] Bhattacharyya, A. (1946). 'On a measure of divergence between two multinomial populations'. *Sankhyā: the Indian Journal of Statistics*, pp. 401–406.

[69] Balci, O. and Sargent, R.G. (1982). 'Validation of multivariate response models using Hotelling's two-sample t^2 test'. *Simulation*, 39(6), pp. 185–192.

[70] Kullback, S. and Leibler, R.A. (1951). 'On information and sufficiency'. *Annals of Mathematical Statistics*, 22(1), pp. 79–86.

[71] Oden, J.T., Prudencio, E.E. and Bauman, P.T. (2013). 'Virtual model validation of complex multiscale systems: Applications to nonlinear elastostatics'. *Computer Methods in Applied Mechanics and Engineering*, 266, pp. 162–184.

[72] Neyman, J. and Pearson, E.S. (1928). 'On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I'. *Biometrika*, 20A(1/2), pp. 175–240.

[73] Neyman, J. and Pearson, E.S. (1933). 'IX. On the problem of the most efficient tests of statistical hypotheses'. *Philosophical Transactions of the Royal Society of London. Series A.*, 231(694-706), pp. 289–337.

[74] Berger, R.L. and Hsu, J.C. (1996). 'Bioequivalence trials, intersection-union tests and equivalence confidence sets'. *Statistical Science*, 11(4), pp. 283–319.

[75] Robinson, A.P. and Froese, R.E. (2004). 'Model validation using equivalence tests'. *Ecological Modelling*, 176(3-4), pp. 349–358.

[76] Rebba, R. and Mahadevan, S. (2008). 'Computational methods for model reliability assessment'. *Reliability Engineering and System Safety*, 93(8), pp. 1197–1207.

[77] Thacker, B.H. and Paez, T.L. (2014). 'A simple probabilistic validation metric for the comparison of uncertain model and test results'. In '16th AIAA Non-Deterministic Approaches Conference', pp. 0121.

[78] Tobler, W.R. (1970). 'A computer movie simulating urban growth in the Detroit region'. *Economic Geography*, 46(sup1), pp. 234–240.

[79] Griffith, D.A. (2005). 'Effective geographic sample size in the presence of spatial autocorrelation'. *Annals of the Association of American Geographers*, 95(4), pp. 740–760.

[80] Myers, D.E. (1989). 'To be or not to be... stationary? That is the question'. *Mathematical Geology*, 21(3), pp. 347–362.

[81] Oliver, M.A. and Webster, R. (2014). 'A tutorial guide to geostatistics: Computing and modelling variograms and kriging'. *CATENA*, 113, pp. 56–69.

[82] Andrianakis, I., Vernon, I.R., McCreesh, N., McKinley, T.J., Oakley, J.E., Nsubuga, R.N., Goldstein, M. and White, R.G. (2015). 'Bayesian history matching of complex infectious disease models using emulation: a tutorial and a case study on HIV in Uganda'. *PLoS Computational Biology*, 11(1), pp. 1003968.

[83] Clifford, P. and Richardson, S. (1985). 'Testing the association between two spatial processes'. *Statistics and Decisions*, 2(Supp. issue), pp. 155–160.

[84] Dutilleul, P., Clifford, P., Richardson, S. and Hemon, D. (1993). 'Modifying the t Test for Assessing the Correlation Between Two Spatial Processes'. *Biometrics*, pp. 305–314.

[85] Li, B., Griffith, D.A. and Becker, B. (2016). 'Spatially simplified scatterplots for large raster datasets'. *Geo-spatial Information Science*, 19(2), pp. 81–93.

[86] Levine, R.S., Yorita, K.L., Walsh, M.C. and Reynolds, M.G. (2009). 'A method for statistically comparing spatial distribution maps'. *International Journal of Health Geographics*, 8(7).

[87] Wilson, P.D. (2011). 'Distance-based methods for the analysis of maps produced by species distribution models'. *Methods in Ecology and Evolution*, 2(6), pp. 623–633.

[88] Jones, E.L., Rendell, L., Pirotta, E. and Long, J.A. (2016). 'Novel application of a quantitative spatial comparison tool to species distribution data'. *Ecological Indicators*, 70, pp. 67–76.

[89] Wiederholt, R., Paudel, R., Khare, Y., Davis, S.E., Naja, G.M., Romañach, S., Pearlstine, L. and Van Lent, T. (2019). 'A multi-indicator spatial similarity approach for evaluating ecological restoration scenarios'. *Landscape Ecology*, 34(11), pp. 2557–2574.

[90] Robertson, C., Long, J.A., Nathoo, F.S., Nelson, T.A. and Plouffe, C.C. (2014). 'Assessing quality of spatial models using the structural similarity index and posterior predictive checks'. *Geographical Analysis*, 46(1), pp. 53–74.

[91] Wang, Z., Bovik, A.C., Sheikh, H.R. and Simoncelli, E.P. (2004). 'Image quality assessment: from error visibility to structural similarity'. *IEEE Transactions on Image Processing*, 13(4), pp. 600–612.

[92] Gröning, F., Liu, J., Fagan, M. and O'higgins, P. (2009). 'Validating a voxel-based finite element model of a human mandible using digital speckle pattern interferometry'. *Journal of Biomechanics*, 42(9), pp. 1224–1229.

[93] Dickinson, A., Taylor, A., Ozturk, H. and Browne, M. (2011). 'Experimental validation of a finite element model of the proximal femur using digital image correlation and a composite bone model'. *Journal of Biomechanical Engineering*, 133(1), pp. 014504.

[94] Lomov, S.V., Ivanov, D.S., Verpoest, I., Zako, M., Kurashiki, T., Nakai, H., Molimard, J. and Vautrin, A. (2008). 'Full-field strain measurements for validation of meso-FE analysis of textile composites'. *Composites Part A: Applied Science and Manufacturing*, 39(8), pp. 1218–1231.

[95] Liang, Y., Lee, H., Lim, S., Lin, W., Lee, K. and Wu, C. (2002). 'Proper orthogonal decomposition and its applications—Part I: Theory'. *Journal of Sound and Vibration*, 252(3), pp. 527–544.

[96] Lumley, J.L. (1981). 'Coherent Structures in Turbulence'. In R.E. Mayer (editor), 'Transition and Turbulence', pp. 215–242. Elsevier.

[97] Berkooz, G., Holmes, P. and Lumley, J.L. (1993). 'The proper orthogonal decomposition in the analysis of turbulent flows'. *Annual Review of Fluid Mechanics*, 25(1), pp. 539–575.

[98] Bistrian, D. and Susan-Resiga, R. (2016). 'Weighted proper orthogonal decomposition of the swirling flow exiting the hydraulic turbine runner'. *Applied Mathematical Modelling*, 40(5-6), pp. 4057–4078.

[99] Schmid, P.J. (2010). 'Dynamic mode decomposition of numerical and experimental data'. *Journal of Fluid Mechanics*, 656, pp. 5–28.

[100] Schmid, P.J. (2011). 'Application of the dynamic mode decomposition to experimental data'. *Experiments in Fluids*, 50(4), pp. 1123–1130.

[101] Seena, A. and Sung, H.J. (2011). 'Dynamic mode decomposition of turbulent cavity flows for self-sustained oscillations'. *International Journal of Heat and Fluid Flow*, 32(6), pp. 1098–1110.

[102] Guéniat, F., Pastur, L. and Lusseyran, F. (2014). 'Investigating mode competition and three-dimensional features from two-dimensional velocity fields in an open cavity flow by modal decompositions'. *Physics of Fluids*, 26(8), pp. 085101.

[103] Wang, W., Mottershead, J.E., Sebastian, C.M. and Patterson, E.A. (2011). 'Shape features and finite element model updating from full-field strain data'. *International Journal of Solids and Structures*, 48(11-12), pp. 1644–1657.

[104] Patki, A. and Patterson, E. (2012). 'Decomposing strain maps using Fourier-Zernike shape descriptors'. *Experimental Mechanics*, 52(8), pp. 1137–1149.

[105] Worden, K., Manson, G. and Fieller, N.R. (2000). 'Damage detection using outlier analysis'. *Journal of Sound and Vibration*, 229(3), pp. 647–667.

[106] Kopsaftopoulos, F.P. and Fassois, S.D. (2010). 'Vibration based health monitoring for a lightweight truss structure: experimental assessment of several statistical time series methods'. *Mechanical Systems and Signal Processing*, 24(7), pp. 1977–1997.

[107] Farrar, C., Nishio, M., Hemez, F., Stull, C., Park, G., Cornwell, P., Figueiredo, E., Luscher, D. and Worden, K. (2016). *Feature Extraction for Structural Dynamics Model Validation*. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

[108] Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

[109] Loutas, T., Roulias, D., Pauly, E. and Kostopoulos, V. (2011). 'The combined use of vibration, acoustic emission and oil debris on-line monitoring towards a more effective condition monitoring of rotating machinery'. *Mechanical Systems and Signal Processing*, 25(4), pp. 1339–1352.

[110] Li, L. and Lu, Z. (2018). 'A new method for model validation with multivariate output'. *Reliability Engineering & System Safety*, 169, pp. 579–592.

[111] Liu, Y., Sun, X. and Dinh, N.T. (2019). 'Validation and uncertainty quantification of multiphase-CFD solvers: A data-driven Bayesian framework supported by high-resolution experiments'. *Nuclear Engineering and Design*, 354, pp. 110200.

[112] Wang, Z., Fu, Y., Yang, R.J., Barbat, S. and Chen, W. (2016). 'Validating Dynamic Engineering Models Under Uncertainty'. *Journal of Mechanical Design*, 138(11), pp. 111402.

[113] Dubois, D.J. (1980). *Fuzzy Sets and Systems: Theory and Applications*, volume 144. Academic Press.

[114] Moore, R.E. (1966). *Interval Analysis*. Prentice-Hall.

[115] Ferson, S. and Hajagos, J.G. (2004). 'Arithmetic with uncertain numbers: rigorous and (often) best possible answers'. *Reliability Engineering & System Safety*, 85(1-3), pp. 135–152.

[116] Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press.

[117] Rubinstein, R.Y. and Kroese, D.P. (2016). *Simulation and the Monte Carlo method*. John Wiley & Sons.

[118] Bickel, P.J. and Freedman, D. (1981). 'Asymptotic theory for the bootstrap'. *Annals of Statistics*, 9(6), pp. 1196–1217.

[119] Angus, J.E. (1994). 'Probability integral transform and related results'. *SIAM Review*, 36(4), pp. 652–654.

[120] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

[121] McKay, M.D., Beckman, R.J. and Conover, W.J. (1979). 'A Comparison of Three methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code'. *Technometrics*, 21(1), pp. 239–245.

[122] Mukundan, R., Ong, S.H. and Lee, P.A. (2001). 'Image analysis by Tchebichef moments'. *IEEE Transactions on Image Processing*, 10(9), pp. 1357–1364.

[123] Berke, R.B., Sebastian, C.M., Chona, R., Patterson, E.A. and Lambros, J. (2016). 'High Temperature Vibratory Response of Hastelloy-X: Stereo-DIC Measurements and Image Decomposition Analysis'. *Experimental Mechanics*, 56(2), pp. 231–243.

[124] Van der Vaart, A.W. (2000). *Asymptotic Statistics*, volume 3. Cambridge University Press.

[125] Minasny, B (2020). 'Latin Hypercube Sampling'. URL https://www.mathworks.com/matlabcentral/fileexchange/4352-latin-hypercube-sampling. Accessed 11 January 2020.

[126] North, G.R., Bell, T.L., Cahalan, R.F. and Moeng, F.J. (1982). 'Sampling errors in the estimation of empirical orthogonal functions'. *Monthly Weather Review*, 110(7), pp. 699–706.

[127] Babamoradi, H., van den Berg, F. and Rinnan, Å. (2013). 'Bootstrap based confidence limits in principal component analysis—A case study'. *Chemometrics and Intelligent Laboratory Systems*, 120, pp. 97–105.

[128] Kang, J., Jin, R., Li, X. and Zhang, Y. (2017). 'Block Kriging With Measurement Errors: A Case Study of the Spatial Prediction of Soil Moisture in the Middle Reaches of Heihe River Basin'. *IEEE Geoscience and Remote Sensing Letters*, 14(1), pp. 87–91.

[129] Gaillard, F., Reynaud, T., Thierry, V., Kolodziejczyk, N. and von Schuckmann, K. (2016). 'In situ–based reanalysis of the global ocean temperature and salinity with isas: Variability of the heat content and steric height'. *Journal of Climate*, 29(4), pp. 1305–1323.

[130] Gaillard, F. (2015). 'ISAS-13 temperature and salinity gridded fields'. URL https://doi.org/10.17882/45945.

[131] Jolliffe, I. (2011). *Principal Component Analysis*. Springer.

[132] Marjoram, P., Molitor, J., Plagnol, V. and Tavaré, S. (2003). 'Markov chain Monte Carlo without likelihoods'. *Proceedings of the National Academy of Sciences*, 100(26), pp. 15324–15328.

[133] Haario, H., Saksman, E. and Tamminen, J. (2001). 'An adaptive Metropolis algorithm'. *Bernoulli*, 7(2), pp. 223–242.

[134] Gelman, A. and Rubin, D.B. (1992). 'Inference from iterative simulation using multiple sequences'. *Statistical Science*, 7(4), pp. 457–472.

[135] Brooks, S.P. and Gelman, A. (1998). 'General methods for monitoring convergence of iterative simulations'. *Journal of Computational and Graphical Statistics*, 7(4), pp. 434–455.

[136] Kass, R.E., Carlin, B.P., Gelman, A. and Neal, R.M. (1998). 'Markov Chain Monte Carlo in practice: a roundtable discussion'. *American Statistician*, 52(2), pp. 93–100.

[137] Box, G.E., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

[138] Kruschke, J.K. (2010). 'Bayesian data analysis'. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(5), pp. 658–676.

[139] Hack, E., Lampeas, G. and Patterson, E.A. (2016). 'An evaluation of a protocol for the validation of computational solid mechanics models'. *The Journal of Strain Analysis for Engineering Design*, 51(1), pp. 5–13.

[140] Mottershead, J.E. and Wang, W. (2013). 'Principles of image processing and feature recognition applied to full-field measurements'. In R. Allemang, J. De Clerck, C. Niezrecki and A. Wicks (editors), 'Special Topics in Structural Dynamics, Volume 6', pp. 411–424. Springer.

[141] Christian, W. and Patterson, E. (2019). 'EUCLID software'. URL http://www.experimentalstress.com/euclid/euclid_conditions.htm. Accessed 30 August 2019.

[142] Picchini, U. (2019). 'ABC-MCMC for g-and-k distributions'. URL https://github.com/umbertopicchini/abc_g-and-k. Accessed 7 October 2019.

[143] Roemmich, D., Johnson, G.C., Riser, S., Davis, R., Gilson, J., Owens, W.B., Garzoli, S.L., Schmid, C. and Ignaszewski, M. (2009). 'The Argo Program: Observing the global ocean with profiling floats'. *Oceanography*, 22(2), pp. 34–43.

[144] Taira, K., Brunton, S.L., Dawson, S.T., Rowley, C.W., Colonius, T., McKeon, B.J., Schmidt, O.T., Gordeyev, S., Theofilis, V. and Ukeiley, L.S. (2017). 'Modal analysis of fluid flows: An overview'. *AIAA Journal*, 55(12), pp. 4013–4041.

[145] Mingqiang, Y., Kidiyo, K. and Joseph, R. (2008). 'A survey of shape feature extraction techniques'. *Pattern Recognition*, 15(7), pp. 43–90.

[146] Christensen, W.F. (2011). 'Filtered Kriging for Spatial Data with Heterogeneous Measurement Error Variances'. *Biometrics*, 67(3), pp. 947–957.

[147] García-Soidán, P., Menezes, R. and Rubiños, Ó. (2014). 'Bootstrap approaches for spatial data'. *Stochastic Environmental Research and Risk Assessment*, 28(5), pp. 1207–1219.

[148] Castillo-Páez, S., Fernández-Casal, R. and García-Soidán, P. (2019). 'A nonparametric bootstrap method for spatial data'. *Computational Statistics & Data Analysis*, 137, pp. 1–15.

[149] Cai, T.T., Liang, T. and Zhou, H.H. (2015). 'Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional gaussian distributions'. *Journal of Multivariate Analysis*, 137, pp. 161–172.

[150] Gneiting, T., Stanberry, L.I., Grimit, E.P., Held, L. and Johnson, N.A. (2008). 'Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds'. *Test*, 17(2), pp. 211.

[151] National Oceanic and Atmospheric Administration (2019). 'El Niño & La Niña (El Niño-Southern Oscillation)'. URL https://www.climate.gov/enso. Accessed 7 June 2019.

[152] Dahlman, L. (2019). 'Climate variability: Oceanic Niño Index'. URL https://www.climate.gov/news-features/understanding-climate/climate-variability-oceanic-ni~no-index. Accessed 7 June 2019.

[153] National Oceanic and Atmospheric Administration (2019). 'Cold & Warm Episodes by Season'. URL https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php. Accessed 4 June 2019.

[154] Kullaa, J. (2006). 'Removing non-linear environmental influences from structural features'. In 'Proceedings of the third European Workshop on Structural Health Monitoring, Granada, Spain', pp. 635–642.

[155] Figueiredo, E., Park, G., Figueiras, J., Farrar, C. and Worden, K. (2009). *Structural health monitoring algorithm comparisons using standard data*

sets. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

[156] Wang, D., Diazdelao, F.A., Wang, W., Lin, X., Patterson, E.A. and Mottershead, J.E. (2016). 'Uncertainty quantification in DIC with Kriging regression'. *Optics and Lasers in Engineering*, 78, pp. 182–195.

[157] Ke, X.D., Schreier, H., Sutton, M. and Wang, Y. (2011). 'Error assessment in stereo-based deformation measurements'. *Experimental Mechanics*, 51(4), pp. 423–441.

[158] Ling, Y. and Mahadevan, S. (2013). 'Quantitative model validation techniques: New insights'. *Reliability Engineering and System Safety*, 111, pp. 217–231.

[159] Lin, L.I.K. (1989). 'A concordance correlation coefficient to evaluate reproducibility'. *Biometrics*, pp. 255–268.

[160] Avril, S., Bonnet, M., Bretelle, A.S., Grédiac, M., Hild, F., Ienny, P., Latourte, F., Lemosse, D., Pagano, S., Pagnacco, E. *et al.* (2008). 'Overview of identification methods of mechanical parameters based on full-field measurements'. *Experimental Mechanics*, 48(4), pp. 381.

[161] Gogu, C., Yin, W., Haftka, R., Ifju, P., Molimard, J., Le Riche, R. and Vautrin, A. (2013). 'Bayesian identification of elastic constants in multi-directional laminate from moiré interferometry displacement fields'. *Experimental Mechanics*, 53(4), pp. 635–648.

Appendices

Appendix A

Comparison of validation metrics: 1D numerical examples

The figures corresponding to the results outlined in table 3.1 for the 1D numerical examples of Chapter 3 are shown .

Table A.1: Numerical examples' (univariate) parameter definition.

No.	$\mu_{exp}(\mu\epsilon)$	$\mu_{sim}(\mu\epsilon)$	$\sigma_{exp}(\mu\epsilon)$	$\sigma_{sim}(\mu\epsilon)$	N_{exp}	N_{sim}	area metric ($\mu\epsilon$)	u-pooling
1	150	150	3	3	1000	1000	0.32	0.03
2	150	150	3	3	6	1000	0.97	0.08
3	150	150	3	6	1000	1000	2.75	0.12
4	150	150	3	6	6	1000	3.43	0.15
5	150	150	3	12	1000	1000	7.81	0.18
6	150	150	3	12	6	1000	8.54	0.20
7	150	150	3	24	1000	1000	17.96	0.22
8	150	150	3	24	6	1000	18.80	0.23
9	147	150	3	3	1000	1000	3.27	0.27
10	147	150	3	3	6	1000	3.25	0.30
11	144	150	3	3	1000	1000	6.27	0.43
12	144	150	3	3	6	1000	6.16	0.45
13	142	150	3	3	1000	1000	8.28	0.47
14	142	150	3	3	6	1000	8.03	0.49

μ_{exp}	mean value for the experimental dataset
μ_{sim}	mean value for the simulated dataset
σ_{exp}	standard deviation for the experimental dataset
σ_{sim}	standard deviation for the simulated dataset
N_{exp}	number of experimental measurements
N_{sim}	number of simulation outputs
area metric	area metric calculation result
u-pooling	u-pooling calculation result

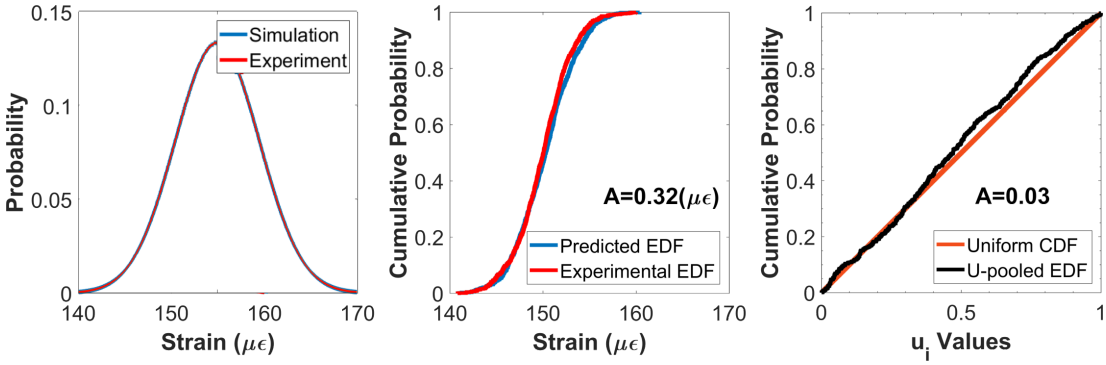

Figure A.1: Test 1: measurements and predictions follow the same distribution. A nonzero validation outcome, as shown by the A value in the figure, arises due to the discrete nature of the distribution functions. $N_{exp} = N_{sim} = 1000$.

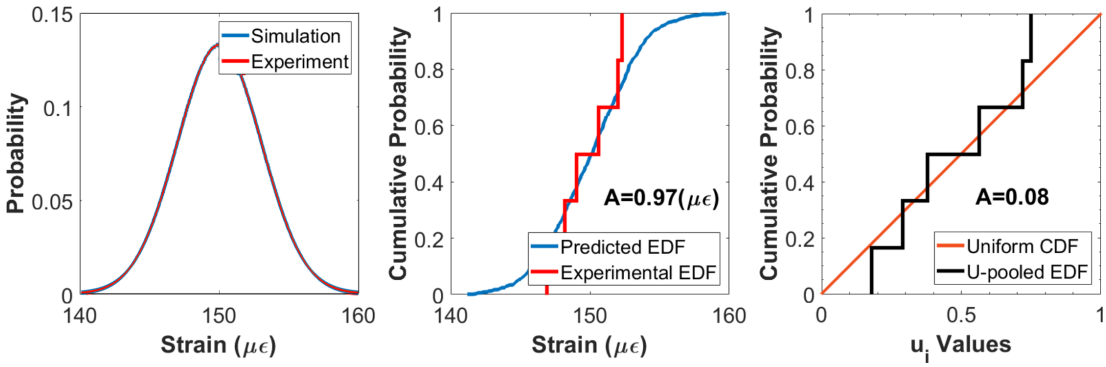

Figure A.2: Example 2: measurements and predictions follow the same distribution. The discrepancy between the two increases as the number of measurements decreases. $N_{exp} = 6, N_{sim} = 1000$.

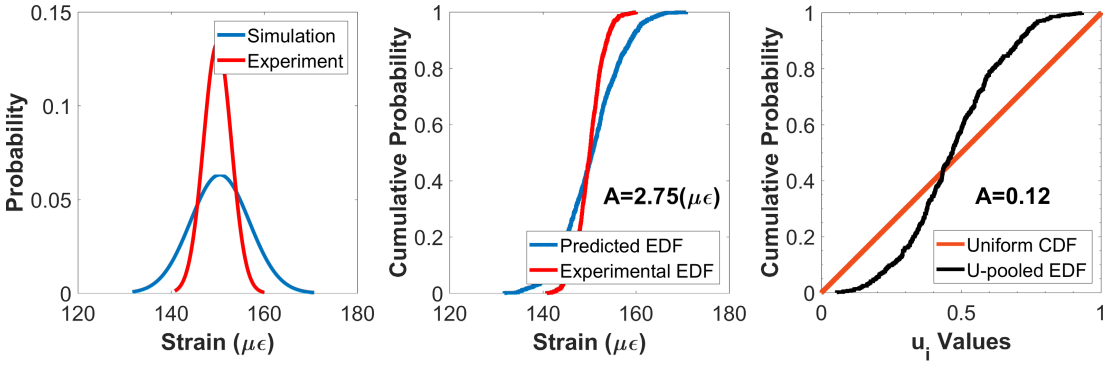

Figure A.3: Example 3: the mean across the two distributions is the same while the standard deviations differ. $\sigma_{exp} = 3\ \mu\epsilon, \sigma_{sim} = 6\ \mu\epsilon$.

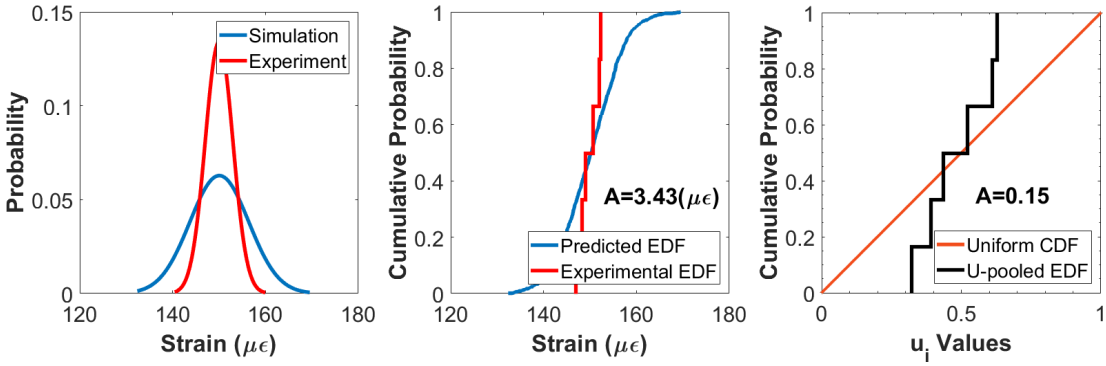

Figure A.4: Example 4: similar to figure A.3 but the number of measurements differs. $N_{exp} = 6, N_{sim} = 1000$.

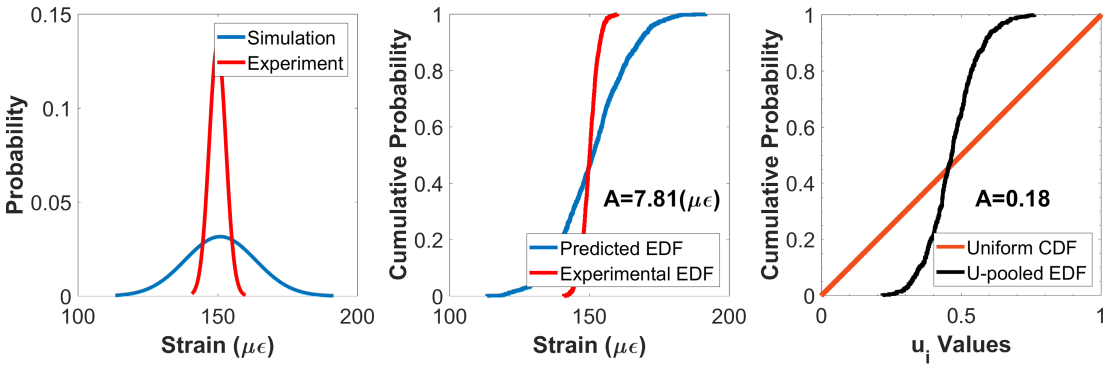

Figure A.5: Example 5: the standard deviation of the simulation outputs is $\sigma_{sim} = 12\ \mu\epsilon$ while that of the measurements is $\sigma_{exp} = 3\ \mu\epsilon$.

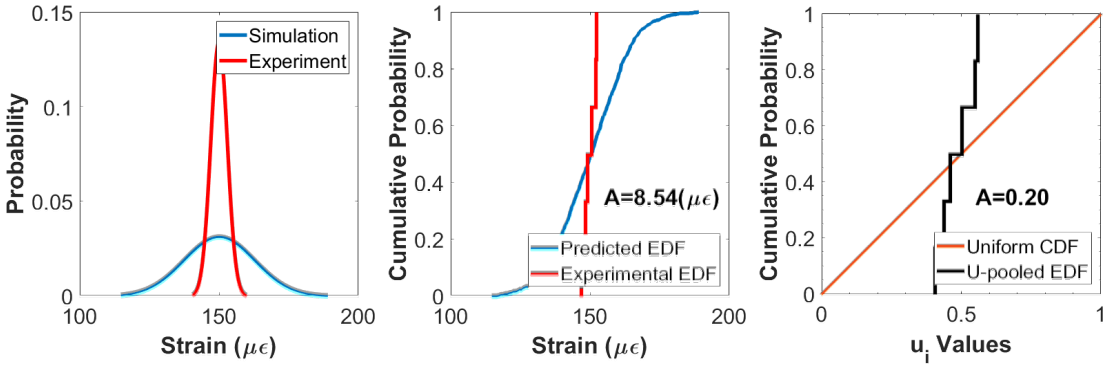

Figure A.6: Example 6: similar to figure A.5 but the number of measurements differs. $N_{exp} = 6, N_{sim} = 1000$.

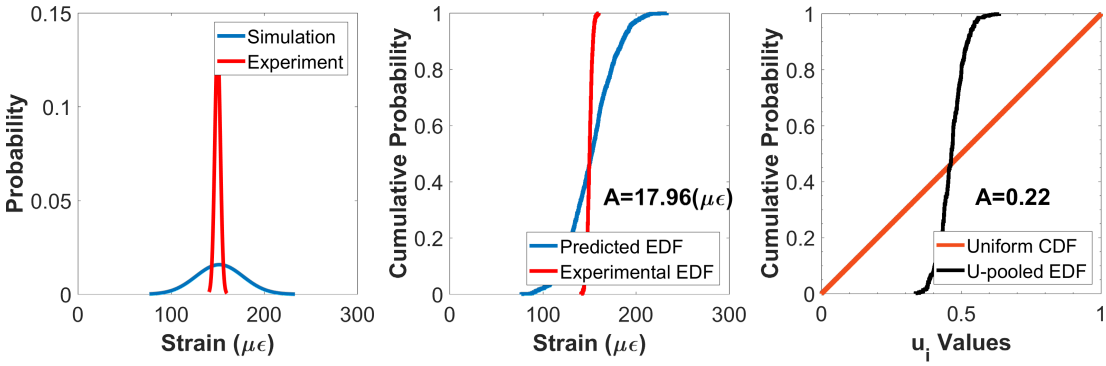

Figure A.7: Example 7: the standard deviation of the simulation outputs is $\sigma_{sim} = 24\ \mu\epsilon$ while that of the measurements is $\sigma_{exp} = 3\ \mu\epsilon$.

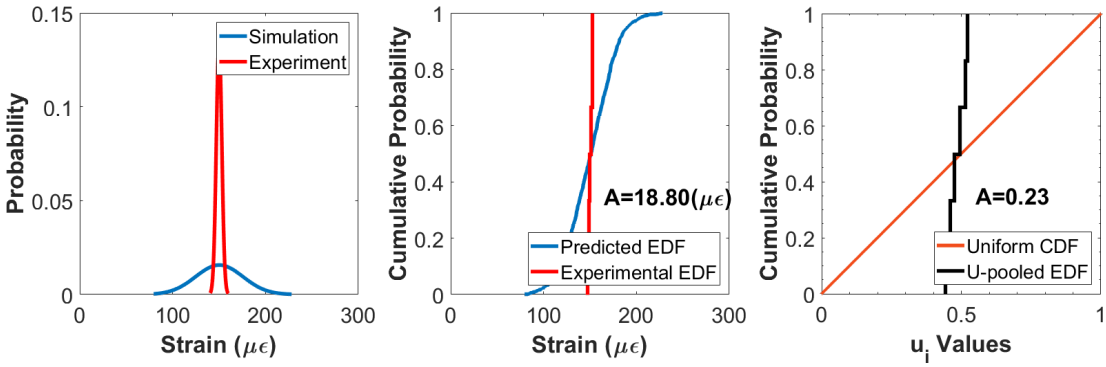

Figure A.8: Example 8: similar to figure A.7 but the number of measurements differs. $N_{exp} = 6, N_{sim} = 1000$.

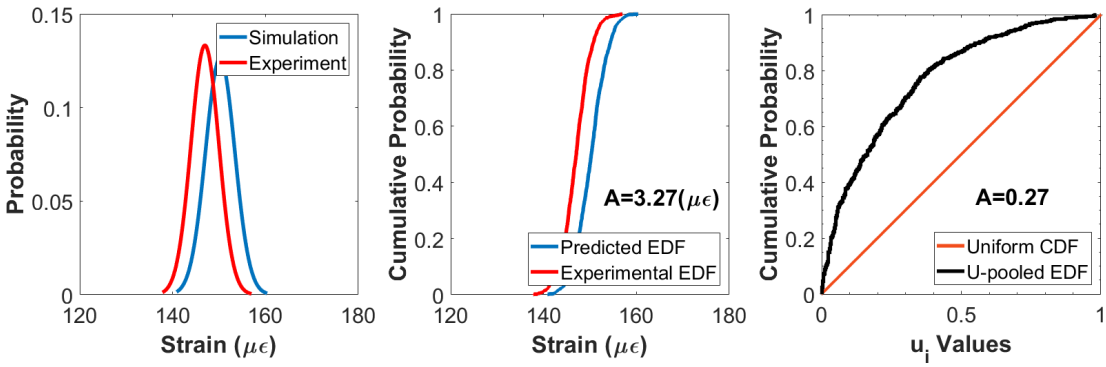

Figure A.9: Example 9: the standard deviations across the two distributions are the same but their means differ. $\mu_{exp} = 147\ \mu\epsilon, \mu_{sim} = 150\ \mu\epsilon$.

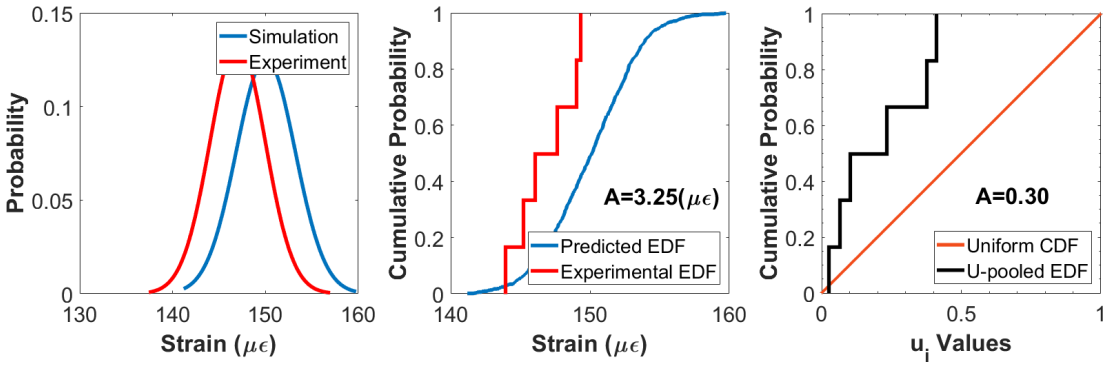

Figure A.10: Example 10: similar to figure A.9 but the number of measurements differs. $N_{exp} = 6, N_{sim} = 1000$.

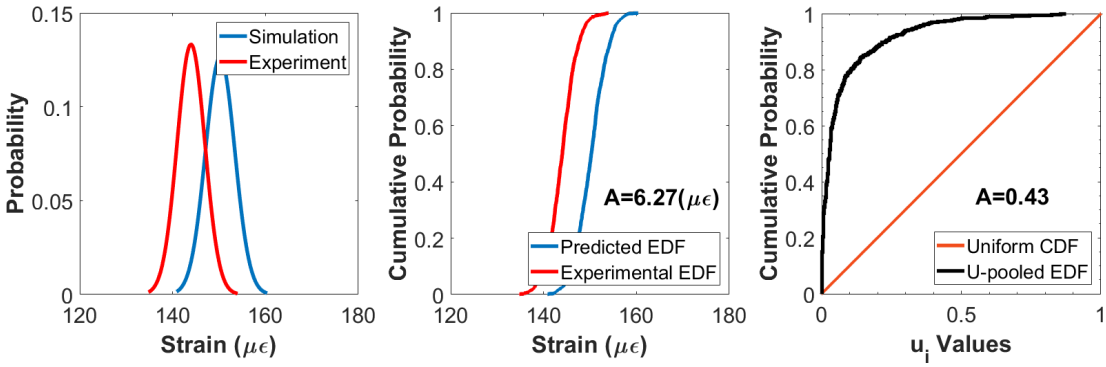

Figure A.11: Example 11: the standard deviations across the two distributions are the same but their means differ. $\mu_{exp} = 144\ \mu\epsilon,\ \mu_{sim} = 150\ \mu\epsilon$.

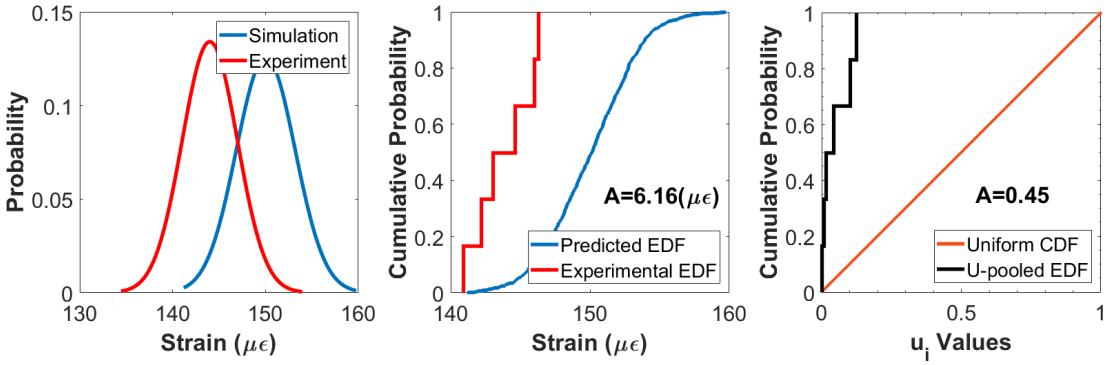

Figure A.12: Example 12: similar to figure A.11 but the number of measurements differs. $N_{exp} = 6, N_{sim} = 1000$.

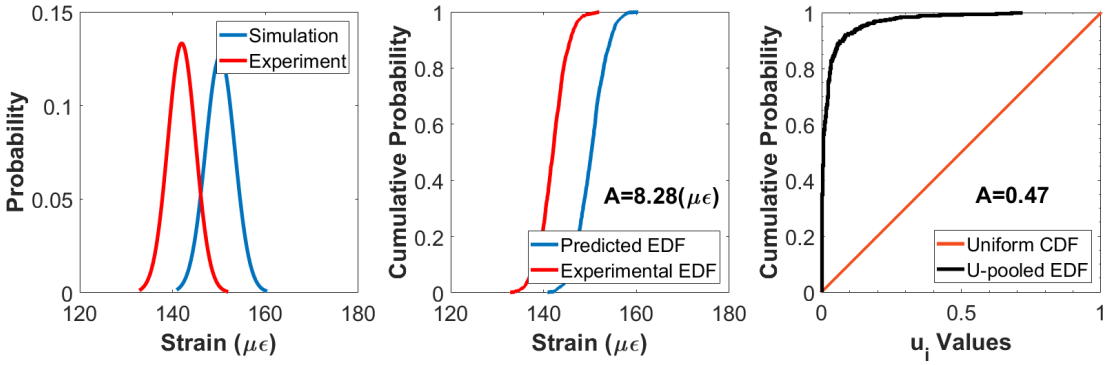

Figure A.13: Example 13: the standard deviations across the two distributions are the same but their means differ. $\mu_{exp} = 142\ \mu\epsilon, \mu_{sim} = 150\ \mu\epsilon$.

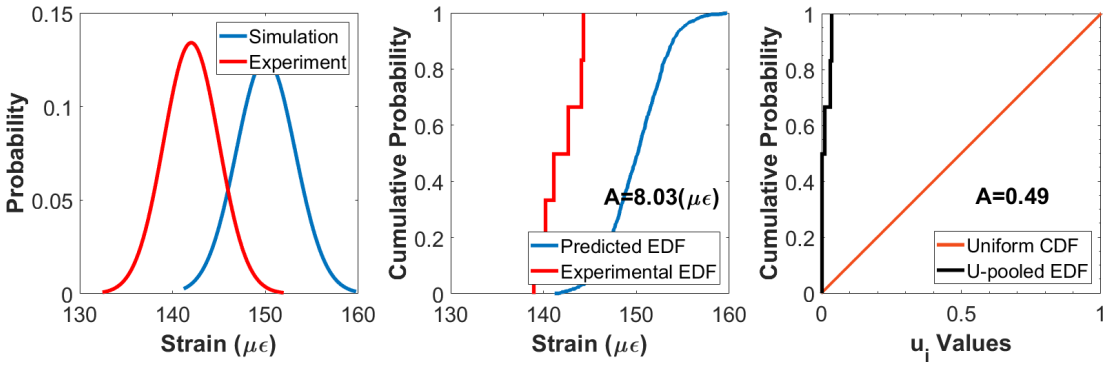

Figure A.14: Example 14: similar to figure A.11 but the number of measurements differs. $N_{exp} = 6, N_{sim} = 1000$.

Appendix B

Comparison of validation metrics: 2D numerical examples

The figures corresponding to the results outlined in table 3.2 for the 2D numerical examples of Chapter 3 are shown.

Table B.1: Parameters and results for the 2-D numerical examples.

	$\mu_{exp_1}(mm)$	$\mu_{exp_2}(\mu\epsilon)$	$\mu_{sim_1}(mm)$	$\mu_{sim_2}(\mu\epsilon)$	$\sigma_{exp_1}(mm)$	$\sigma_{exp_2}(\mu\epsilon)$	$\sigma_{sim_1}(mm)$	$\sigma_{sim_2}(\mu\epsilon)$	ρ_{exp}	ρ_{sim}	MD metric	PIT metric
1	-0.47	1214	-0.47	1214	0.0160	36	0.0160	36	-0.59	-0.59	0.24	0.05
2	-0.46	1244	-0.47	1214	0.0160	36	0.0160	36	-0.59	-0.59	9.31	0.44
3	-0.46	1244	-0.47	1214	0.0320	36	0.0160	36	-0.59	-0.59	9.17	0.30
4	-0.46	1244	-0.47	1214	0.0160	36	0.0160	36	0	-0.59	9.12	0.44
5	-0.48	1184	-0.47	1214	0.0160	36	0.0160	36	-0.59	-0.59	9.50	0.15
6	-0.48	1184	-0.47	1214	0.0160	36	0.0160	36	-0.80	-0.59	9.43	0.15
7	-0.48	914	-0.47	1214	0.0160	36	0.0160	36	-0.59	-0.95	28.40	0.06
8	-0.47	1814	-0.47	1214	0.0160	36	0.0160	36	-0.59	-0.59	19.52	0.44
9	-0.47	614	-0.47	1214	0.0160	36	0.0160	36	-0.59	-0.59	19.69	0.15

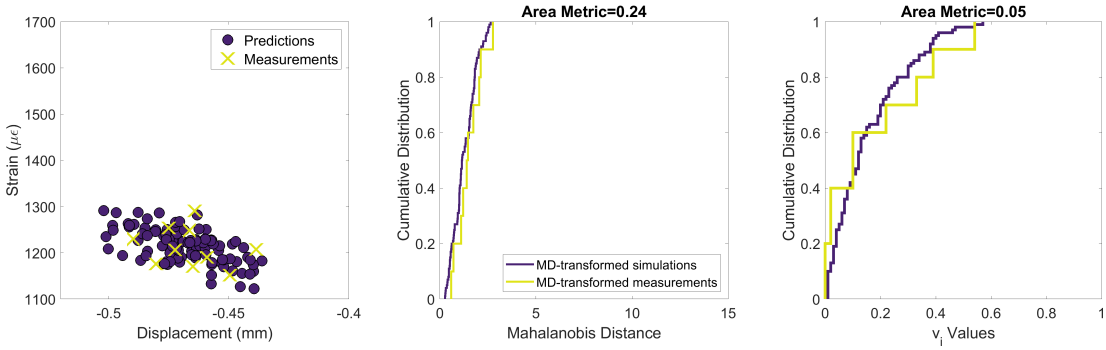

Figure B.1: Example 1: the means, standard deviations and correlations are the same in both datasets. The simulated and measured responses are shown on the left. In the middle, the Mahalanobis-based transformation of the measurements and simulations is portrayed along with the result of their comparison on the title. On the right, the PIT transformation of the simulations is shown in purple along with v-transformed measuremets.

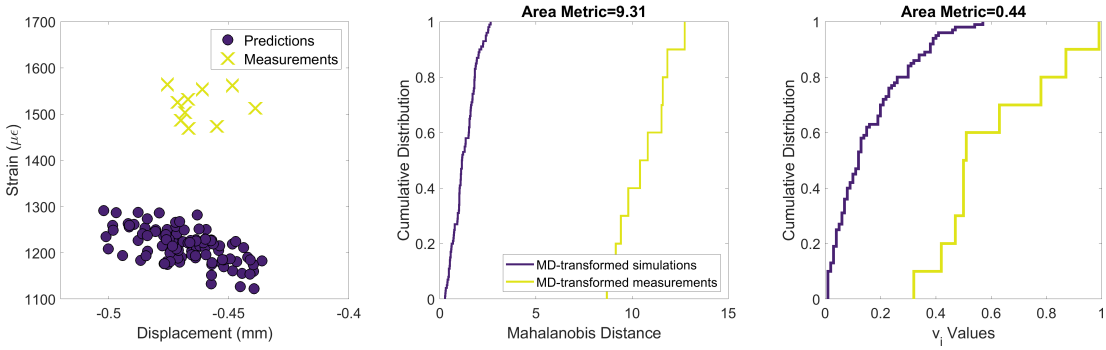

Figure B.2: Example 2: $\mu_{exp_1} = -0.46$ mm $\mu_{exp_2} = 1244$ $\mu\epsilon$, $\mu_{sim_1} = -0.47$ mm $\mu_{sim_2} = 1214$ $\mu\epsilon$. The standard deviations and correlations are equal across the distributions.

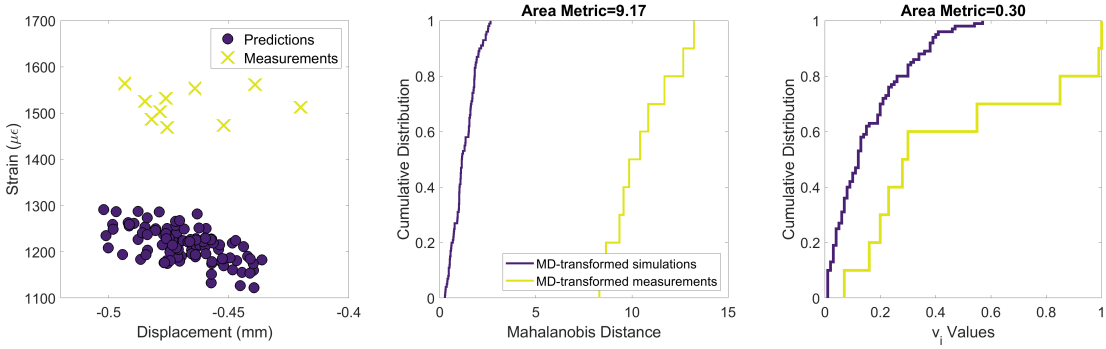

Figure B.3: Example 3: $\mu_{exp_1} = -0.46$ mm $\mu_{exp_2} = 1244$ $\mu\epsilon$, $\mu_{sim_1} = -0.47$ mm $\mu_{sim_2} = 1214$ $\mu\epsilon$ as in the previous example. However, $\sigma_{exp_1} = 0.0320$ mm and $\sigma_{sim_1} = 0.0160$ mm.

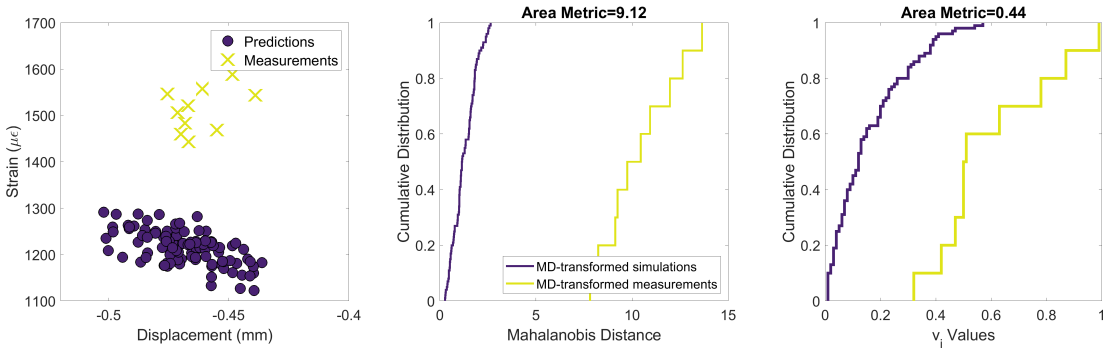

Figure B.4: Example 4: $\mu_{exp_1} = -0.46$ mm $\mu_{exp_2} = 1244$ $\mu\epsilon$, $\mu_{sim_1} = -0.47$ mm $\mu_{sim_2} = 1214$ $\mu\epsilon$ as in the previous example. Standard deviations are equal. $\rho_{exp} = 0$, $\rho_{sim} = -0.59$.

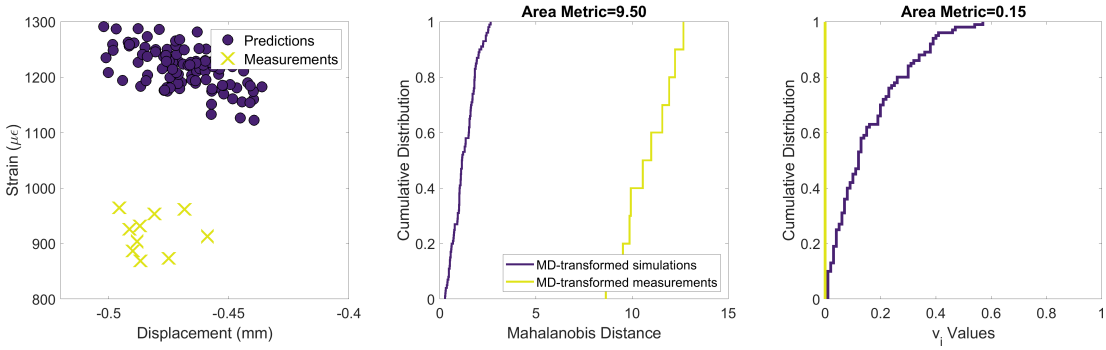

Figure B.5: Example 5: $\mu_{exp_1} = -0.48$ mm $\mu_{exp_2} = 1184$ $\mu\epsilon$, $\mu_{sim_1} = -0.47$ mm $\mu_{sim_2} = 1214$ $\mu\epsilon$. The standard deviations and correlations are equal across the distributions.

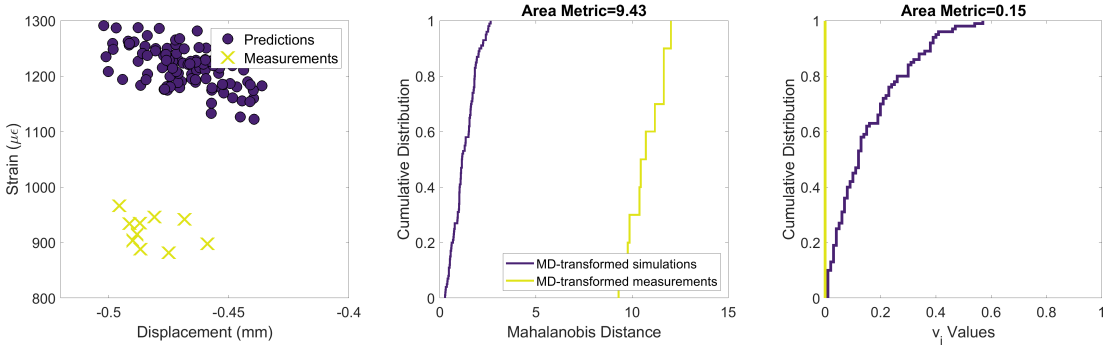

Figure B.6: Example 6: $\mu_{exp_1} = -0.48$ mm $\mu_{exp_2} = 1184$ $\mu\epsilon$, $\mu_{sim_1} = -0.47$ mm $\mu_{sim_2} = 1214$ $\mu\epsilon$ as in the previous example. However, $\rho_{exp} = -0.8$ and $\rho_{sim} = -0.59$.

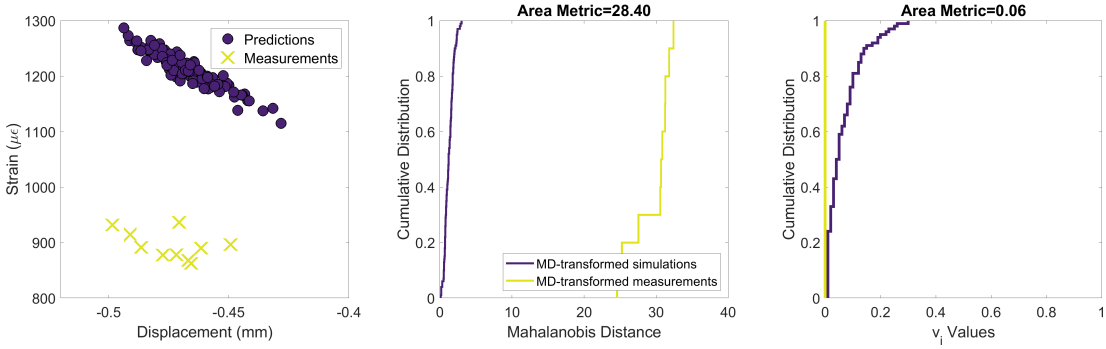

Figure B.7: Example 7: $\mu_{exp_1} = -0.48$ mm $\mu_{exp_2} = 914$ $\mu\epsilon$, $\mu_{sim_1} = -0.47$ mm $\mu_{sim_2} = 1214$ $\mu\epsilon$. There is also a difference in the correlation coefficients. $\rho_{exp} = -0.59$, $\rho_{sim} = -0.95$

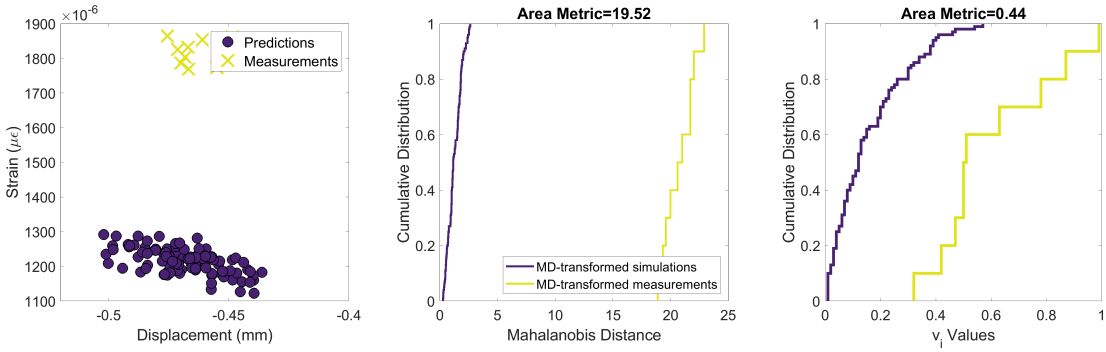

Figure B.8: Example 8: $\mu_{exp_1} = -0.47$ mm $\mu_{exp_2} = 1814$ $\mu\epsilon$, $\mu_{sim_1} = -0.47$ mm $\mu_{sim_2} = 1214$ $\mu\epsilon$. The standard deviations and correlations are equal across the distributions.

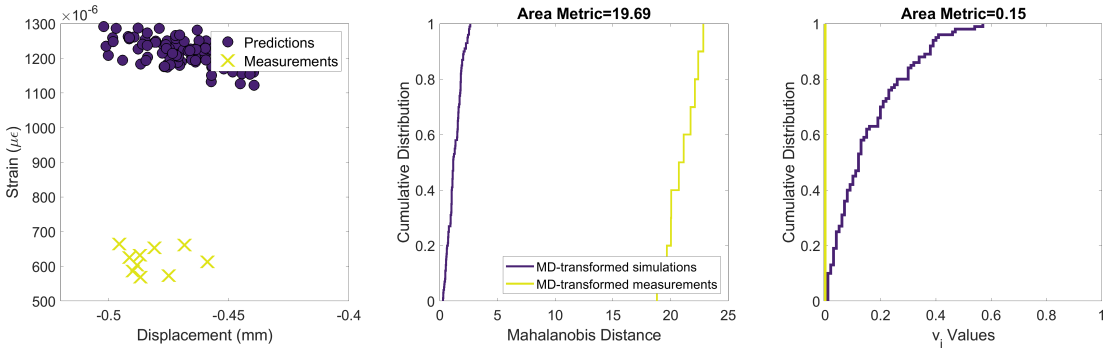

Figure B.9: Example 9: $\mu_{exp_1} = -0.47$ mm $\mu_{exp_2} = 614$ $\mu\epsilon$, $\mu_{sim_1} = -0.47$ mm $\mu_{sim_2} = 1214$ $\mu\epsilon$. The standard deviations and correlations are equal across the distributions.